DOCUMENT RESUME

ED 194 718	CE 026 853

AUTHOR	Cook, Daniel W.: Cooper, Paul G.
TITLE	Fundamentals of Evaluation Research in Vocational Rehabilitation.
INSTITUTION	Arkansas Univ., Fayetteville. Rehabilitation Research and Training Center.
SPONS AGENCY	Rehabilitation Services Administration (DHEW), Washington, D.C. Office of Human Development.
PUB DATE	Mar 76
GRANT	RSA-16-P-56812
NOTE	165p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Annotated Bibliographies: Data Analysis: *Evaluation Methods: Evaluators: Learning Modules: Measurement: Programed Instructional Materials: *Program Evaluation: Rehabilitation Programs: Reliability: *Research Methodology: Sampling: Statistical Analysis: Validity: *Vocational Rehabilitation

ABSTRACT

This training text is designed to introduce program evaluation concepts and terminology to beginning program evaluators for state-federal vocational rehabilitation programs. Taking a research approach, the text covers the principles underlying program evaluation, and is geared to give the reader a basic understanding of how these principles can be applied to the evaluation of human service programs. Designed as a self instructional guide, the text requires little mathematical sophistication on the part of the learner. It is divided into four units, each subdivided into modules which include review exercises. Unit 1 covers general approaches to, and the various issues concerning, program evaluation. Unite 2 deals with sampling concepts: unit 3 with measurement, reliability, validity, and basic psychometric standards: and unit 4 introduces evaluation designs and data analysis. A 77 item annotated bibliography completes the booklet. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194718

Fundamentals of Evaluation Research in Vocational Rehabilitation

Daniel W Cook — Paul G. Cooper

CE 026 853

Arkansas Rehabilitation Research and Training Center
University of Arkansas, Fayetteville

March, 1978

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Daniel Cook
Paul G. Cooper

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

2

All programs administered and the services provided by the Arkansas Rehabilitation Research and Training Center are rendered on a nondiscriminatory basis without regard to handicap, race, color, creed, sex, or national origin in compliance with the Rehabilitation Act of 1973 and Title VI of the Civil Rights Act of 1964. All applicants for program participation and/or services have a right to file complaints and to appeal according to regulations governing this principle.

This study was supported in part by a research and training center grant (16-P-56812, RT-13) from the Rehabilitation Services Administration, Office of Human Development, Department of Health, Education, and Welfare.

Copyright(c) 1978 Daniel W.
Cook and Paul G. Cooper

Errata

p. 3	Printing cost prohibited publication in a loose leaf binder
p. 83	Figure 3 is an error, 34% should be in the space between the mean and ± 1sd, 13.59% should be in the space between ± 1sd and ± 2sd, and 2.14% between ± 2sd and ± 3sd

Preface

The purpose of this publication is to introduce program evaluation concepts and terminology to beginning program evaluators. The emphases and applications are in the context of the State-Federal vocational rehabilitation program. Evaluators in other human service disciplines should also find principles covered herein useful.

This training text is designed as an introductory self-instructional guide to program evaluation. The text may also be used to supplement other materials in a first level graduate course in program evaluation. It requires little mathematical sophistication on the part of the reader. The text is designed in a loose leaf format consisting of four separate units identified by quick reference tabs. Each unit is subdivided into short modules dealing with unit relevant topics. Each module is followed by review exercises. The loose leaf binder facilitates the addition of other program evaluation materials so that the text can be expanded into a reference manual.

Our bias is toward a research approach to program evaluation, and the reader will not find a cook book approach to the current rehabilitation program evaluation standards. However, resources for such guidelines are included in the annotated bibliography. Rather, the text covers an introduction to the principles underlying program evaluation and is geared to give the reader a basic understanding of how these principles can be applied to the evaluation of human service programs. Unit One covers general approaches to, and the various issues concerning, program evaluation. Unit Two deals with sampling concepts; Unit Three with measurement, reliability, validity, and basic psychometric statistics; and Unit Four introduces evaluation designs and data analysis.

We deeply appreciate the helpful comments and editorial assistance of Reed Greenwood, the proofreading ability of Mary Blakemore and Bill Edrington, the typing and manuscript preparation of Sharon Stell and Arlene Maddox, and the layout and design work of David Sigman. The support of the Rehabilitation Service Administration and Vernon Glenn, Director of the Arkansas Rehabilitation Research and Training Center is gratefully acknowledged. Order of authorship is alphabetical; no implication of senority is intended. Accordingly, both authors accept equal responsibility for the accuracy of the material.

January, 1978

DWC
PGC

Unit I

Issues in Program Evaluation

Table of Contents

Module **Page**

Introduction 7
Objective: To discuss the reasons for and general approaches to program evaluation

The Structure of Program Evaluation 13
Objective: To discuss the goals and objectives of program evaluation within the context of the state-federal vocational rehabilitation system

Program Evaluation in a Political Context 19
Objective: To discuss sources of potential resistance to efforts of program evaluators and suggest approaches for avoiding such resistance

Program Evaluation and the Scientific Method 25
Objective: To introduce the scientific method as it applies to the problems of program evaluation

Unit One

Module One

Introduction

Everyone most likely can define what evaluation means. English and English (1958) in their comprehensive dictionary of psychological terms define evaluation as:

> 1. determining the relative importance of something in terms of a standard. 2.[educ.] measurement of the attainment of educational goals, however defined — usually included is a study of relative effectiveness of regulated conditions in furthering or hindering such attainment ...

Man has been evaluating since the beginning, "Did the arrow hit the rabbit?", and evaluation continues to be one of his most pervasive activities, "Is this automobile better than another?" "Will this ad sell?", "Which decision is best?" Evaluation is universal and implies that some **information** is needed to be applied to some **standard** so that a **decision** can be reached. As such evaluation means that information is collected usually through some measurement process, interpreted in comparison to other information or normative standards and the interpretation used to facilitate decision making.

Program evaluation is a relatively new concept in terms of social programs. Prior to 1960 it is difficult to find even one reference to "program evaluation." Being such a new discipline, there are many, perhaps several dozen, definitions of program evaluation. Trantow (1970) has given a representative definition of program evaluation:

> Evaluation is essentially an effort to determine what changes occur as the result of a planned program by comparing actual changes (results) with desired changes (stated goals) and by identifying the degree to which the activity (planned program) is responsible for the changes (p. 3).

Present day program evaluation philosophy and techniques can be traced to several sources, a major one being the development of applied research. During the 1930's and 1940's Kurt Lewin and others applied the methods of traditional scientific research to alleviate pressing social problems. Lewin called this type of inquiry action research and sought to use the environment and the people in it as a laboratory. Lewin's studies on persuading people to eat cheaper cuts of meat during World War II are an example of applied action research. Other developments which have influenced present day program evaluation methods include improved management monitoring techniques largely developed in the Department of Defense during the 1950's and 1960's and technological advances, particularly the development of high speed computers which store and tabulate vast amounts of information.

Perhaps the greatest influence in spurring the development and conduct of program evaluation was, and is, a political awareness of dwindling resources and the need for public accounting of the vast sums of money spent on various social programs. The current mood of "less is more" in government seems certain to carry over into the 1980's and programs such as vocational rehabilitation will continue to be held accountable.

With the advent of program evaluation as an accepted part of the vocational rehabilitation system, a new and different type of person has joined the ranks of vocational rehabilitation agency personnel — the PROGRAM EVALUATION SPECIALIST. Too often, the typical program evaluation specialists, as seen by some other vocational rehabilitation personnel, is 5 feet 10 inches tall and slightly overweight from spending too much time sitting and thinking. Invariably, they are seen as being assisted by a large staff of secretaries who do most of the real work with their job consisting of various and sundry duties including everything from creating work for others to spying on rehabilitation counselors.

The precise evolutionary origins of this new species is currently unknown, however, there seem to be two separate subspecies. One subspecies evolved from within the ranks of the rehabilitation agency. These turncoats who used to be "one of us" have fallen under the influence of "the administration" and "the feds." The second subspecies seem to be "hired guns" from outside the organization. These outsiders are considered the most dangerous since they don't even speak the same language as most rehabilitation professionals.

Despite the existing views held by many agency personnel, program evaluation specialists have an important contribution to make to agency effectiveness. In fact, it is through their efforts that the effectiveness of the program can be determined. Not only that, federal regulations require their presence.

Reasons for Program Evaluation

There are several general reasons for conducting program evaluations. Archibald (cited in Trantow, 1970) has listed five.

1. **Vindication.** Program evaluation is conducted to collect facts illustrating that the program is worthwhile. Evaluation is seen as justifying (hopefully!) the program's existence and is used to demonstrate the need for program continuation.
2. **Salesmanship.** To expand the program through the spread of effective program component parts to other areas where there are similar needs.
3. **Verification.** The determination of program impact, effectiveness and efficiency in relation to program goals and relative to other similar programs. The idea is to discover what the program is really doing.
4. **Improvement.** Analysis of selected information is used to specify how the program can be made more effective and efficient. Program weaknesses are minimized, program strengths maximized.
5. **Understanding.** Program evaluation seeks to understand how the program functions. Specific things that lead to program success or failure are analyzed in order to comprehend the strengths and weaknesses of the program. One searches for the principles by which the program functions.

Types of Program Evaluation

Given a general definition of program evaluation and reasons why it should be conducted one turns to how program evaluation should be conducted. There are two broad approaches to program evaluation, the **goal attainment model** and the **systems analysis method**. Within the two approaches there are several ways assessment can be conducted. Hefferin and Katz (1971) pointed out that assessment can be statistical, clinical, social, psychological or vocational in nature.

There are two general approaches to program evaluation, systems analysis and goal attainment.

Within vocational rehabilitation programs much program evaluation utilizes the goal attainment model, where program worth is established if a program meets its stated goals. In other words, did the program accomplish what it said it would do? The logic behind this approach is readily apparent, the program is evaluated in terms of its effectiveness in meeting predetermined objectives. Unfortunately in practice the process is never so clear cut. Goals are often value judgements, and people can honestly disagree as to the value of a goal. Then too, goals may be explicit (e.g., number of case closures in a specified time) or implicit (maintaining an organizational power structure). One of the biggest problems facing an evaluator is to clearly specify program goals. Goals may be established at different organizational levels, be quite different at each level, and may even be incompatible. For example, higher order goals established outside the program may focus on administrative concerns and emphasize program justification or cost accountability. Such program goals have obvious implications for program survival. On the other hand, goals established at the program level might emphasize program improvement and focus on client outcomes. The actual assessment of goal attainment may involve relative comparisons to establish norms or comparisons to absolute standards such as compliance with regulations. Using the goal attainment approach the evaluator often must refine general statements of purpose (goals) in order to measure actual program impact.

In vocational rehabilitation there is a great deal of emphasis on program evaluation. The Rehabilitation Act of 1973, specified that "the Secretary shall measure and evaluate the impact of all programs authorized by this Act, in order to determine their effectiveness in achieving stated goals in general, and in relation to their cost, their impact on related programs, and their structure and mechanisms for delivery of services including, where appropriate, comparisons with appropriate control groups composed of persons who have not participated in such programs." In this case, program evaluation is goal oriented, the assessment is statistical (cost data, client movement, etc.) and the established goals carry the force of regulations.

The goal attainment model focuses on program effectiveness or outcome. The **systems analysis** approach analyzes functional organizational units to arrive at a judgment of program worth. Using this approach the effectiveness of the organization or program is not judged solely relative to program goals but in terms of the functional relationship of the program's organizational units. Bennett and Weisinger (1974, p. 4) pointed out that "systems approaches seek to evaluate programs by describing the organizational structure of a program, the stated functions of a program, and the process by which that organizational structure with those functions accomplishes the program objectives." The systems approach uses the entire organizational structure and the subunits within that structure in the evaluation scheme. To quote Etzioni (1960, p. 262), the systems analysis approach asks the question "Under given conditions, how close does the organizational allocation of resources approach an optimum distribution?" As might be imagined, extremely large amounts of data are generated through the systems approach. The popularity of this approach has increased with the availability of high speed computers.

Within the systems approach, one is concerned with operations research, which in turn is concerned with organizational activity. On a simplistic-level, operations research examines whether or not agency operations have met agency regulations. Audits and case reviews to assess the degree agency operations meet federal guidelines is an example of operations research. A more sophisticated approach to operations research utilizes cost-benefit and cost-effectiveness analyses as analytic tools. Basically, operations research makes use of program inputs and outputs to evaluate program planning and policy.

On a highly sophisticated level, system analysis considers the entire system and activities within the system to predict probable outcomes under certain conditions. If a set of conditions is known; for example, staffing patterns and allocation of resources, then statistical techniques can be used to select the best administrative action from among a set of alternative actions; e.g., to either minimize inputs (costs) or maximize outputs ("successful" rehabilitation case closures). Other analytical techniques permit model building and simulation studies to facilitate program planning. System analysis is production and management oriented. It is often useful in planning and allocating resources. It is less useful when program inputs and outputs are intangible such as staff morale and client happiness.

Unit One

Module One

Review

1. The three things needed to complete an evaluation are some ________________
applied to some ________________
in order to make a ________________.

2. Three influences that have affected present day program evaluation are:

   1.

   2.

   3.

3. Give your own definition of program evaluation:

4. What are the five reasons Archibald has suggested for conducting program evaluations? Under which reason would you be most likely to conduct an evaluation?

   1.

   2.

   3.

   4.

   5.

5. What are the two general approaches to program evaluation?

6. List some advantages and disadvantages to the goal attainment model of program evaluation.

advantages **disadvantages**

7. List some advantages and disadvantages to the systems analysis approach to program evaluation.

advantages **disadvantages**

References

Bennett, E. C., & Weisinger, M. **Program Evaluation: A Resource Handbook for Vocational Rehabilitation**. New York: ICD Rehabilitation and Research Center, 1974.

English, H., & English, A. **A Comprehensive Dictionary of Psychological and Psychoanalytical Terms.** New York: Longmans, Green & Co., 1958.

Etzioni, A. Two approaches to organizational analysis: a critique and a suggestion. **Administrative Science Quarterly**, 1960, **5**, 257-278.

Hefferin, E., & Katz, A. Issues and orientations in the evaluation of rehabilitation programs. **Rehabilitation Literature**, 1971, **32**, 66-74.

Trantow, D. An introduction to evaluation: Program effectiveness and community needs. **Rehabilitation Literature**, 1970, **31**, 2-9.

Unit One

Module Two

The Structure of Program Evaluation

No matter which program evaluation model one follows, the goal attainment model or the systems analysis approach, the evaluator must link assessment procedures to program planning. In other words the mission of the program and underlying program goals and objectives serve as guides for program assessment procedures. Program mission, goals and objectives are often established in a political context and can lack specificity as to what the program is to accomplish. Human service programs can and do come into existence because of a clear need, but they may have such vague objectives that program success is difficult to measure. It is for this reason that the program evaluator should be involved in the program planning process.

Program Mission

The most general statement of a human service program's purpose is found in its mission statement. The mission statement is comprehensive in that it spells out the purpose for the continuing existence of the program. Examples of mission statements are:

> The mission of the state-federal vocational rehabilitation program is to provide services and resources whereby vocationally disabled persons can obtain earned income and maximum functional independence. The mission of the comprehensive rehabilitation center is to provide comprehensive medical, social, and vocational services whereby the most severely disabled can be restored to maximum vocational functioning.

Program Goals

Program goals follow from and relate to the program mission statement. If vocational rehabilitation has the mission of providing services and resources to the disabled, program goals specify the direction of the program and the conditions necessary for the program to fulfill its mission. One of the goals underlying the vocational rehabilitation program is that evaluation activities will be conducted to assure that the program is addressing its mission. The goal of such an evaluation might be:

1. To provide a method of measuring vocational rehabilitation effectiveness.
2. To provide procedures for measuring state agency efficiency.
3. To provide feedback to program administrators on the effectiveness and efficiency of the program.

Program goals are more specific than program mission statements. Like mission statements they refer to intent and directionality and are not time bound. Unlike mission statements they can change periodically and are based on changing program priorities.

Program Objectives

Program objectives specify how and when a program reaches its goals. They are concrete, time-limited and specific. By carefully stating program objectives the evaluator can determine when the program has achieved its stated goals. Program objectives are by their very nature assessment related. Objectives should be **observable** within a definite **time period** under specified conditions in **measureable ways**. Examples of program objectives are:

1. Within the next 12 months the state agency will increase vocational rehabilitation closures by six percent.
2. By June 1st a comprehensive evaluation plan will be completed.
3. By September 1st rehabilitation clients exposed to XYZ training will acquire good work-related attitudes.

Program objectives are more specific than either program goals or program mission statements. Program goals and objectives define a program.

Program mission statements, goals and objectives all define the purpose of the program, indeed the reason for its existence. In evaluating a program, i.e., obtaining information by which program worth can be judged and program decisions made, the evaluator is concerned with evaluation strategies, tactics, and criteria.

Evaluation Criteria, Strategies and Tactics

Evaluation strategy refers to the broad method by which the program will be evaluated, for example, using the goal attainment model as the method of evaluation. Tactics are the specific techniques or tools used to evaluate program objectives. A specific statistical technique such as correlation analysis is an example of an evaluation tactic. The bottom line of a program evaluation effort is the criterion or standard upon which the evaluation is based. Program objectives state the specifics of what, when, and where a program goal will be accomplished. Criteria specify how we know a program objective has been fulfilled and a program goal met. Without a goal related criterion such a judgement is not possible. Any criterion must be valid (a true measure of program goals), reliable and practical. Generally speaking the more objective and quantifiable the criterion the better it is.

One of the problems evaluators often face is that program objectives are stated in ways that are very difficult to measure and are open to many interpretations. From an evaluators point of view the program objective "by September 1st rehabilitation clients will acquire good work-related attitudes" is vague and unclear. Presumably the program has some means to improve attitudes but "attitudes" is unclear and "good" is open to many interpretations. A better program objective can be stated by attending to the criterion used to assess "work-related attitudes," such as "by September 1st rehabilitation clients enrolled in XYZ training will acquire good work-related attitudes as measured by the Test of Work Attitudes."

The Test of Work Attitudes becomes the criterion or standard for measuring achievement of the program objective. Defining program objectives in terms of a valid and reliable measurement device greatly facilitates measurement of goal attainment. There are, of course, program objectives that do not lend themselves to such direct measurement. In this case program evaluators can use the procedure of **operationally defining** program outcomes. Using this technique a test of program goal attainment consists of defining program objectives in terms of the operations or behaviors thought to be representative of the objective. In other words a term such as work attitude is defined by agreed upon and directly observable behaviors thought to reflect work attitude. Thus the program objective can be written as:

> By September 1st rehabilitation clients enrolled in XYZ training will show a 10% increase in work shop attendance, and a 20% decrease in work shop discipline problems.

The above objective illustrates another dimension of an evaluation criterion. Criteria do not exist in a vacuum, they ultimately relate to some value judgment. As an aid to making a judgment the criterion must be interpretable. The level of criterion behaviors achieved by a program (i.e., the extent the program meets its goals) can be judged relative to some established standard or in an absolute sense. In the above example the criterion is absolute, clients are expected to show a 10% increase in attendance and a 20% decrease in disciplinary problems. Anything less would represent program failure.

Specifying an absolute level of expected outcome precisely defines what is meant by "good" work-related attitudes. On the other hand goal achievement might be judged in a relative manner. That is, program success would be judged relative to previously established attendance rates, and the degree of program success (or failure) judged via the observed post-training attendance rates versus pretraining rates. In reality absolute criterion levels are relative in nature. Absolute levels may not, however, clearly indicate what they are relative to. The important factor is that a level of criterion behavior gains meaning only in comparison to something — a value judgment, empirical evidence, or criterion levels in similar programs.

Program evaluation studies can emphasize different criteria in reaching a conclusion as to a program's worth. The most logical criterion of worth is the program's **effectiveness** or performance in meeting stated goals. The emphasis here is on impact or consequences. A program which has as its goals various operations may be able to show that it was effective in completing those operations. However, unless there is some measure of output or reason for the program's existence beyond mere performance, the evaluation will be of little ultimate value. Other criteria of program worth give attention to the process underlying the ultimate outcomes. One process related criterion is **effort** which is a measure of program inputs or resources expended. Closely related to effort is **efficiency** which Bennett and Weisinger (1974, p. 4) define as "a rate that is produced when the effort expended by a program is compared to the program's effectiveness." Efficiency is different from effectiveness. For example, Nazi Germany was seen, in an industrial/organizational sense, as being very efficient, but from their point of view, the ultimate outcome was not very effective. Effort and efficiency are important to consider in a **formative** program evaluation. Here it is important that program administrators formulate the most efficient ways to meet their stated goals. A formative program evaluation makes use of evaluation data in developing the program. Formative evaluations are ongoing in nature. In a **summative** program evaluation, the emphasis is on program effectiveness or outcome and the adequacy of program performance or quality. Summative evaluations are usually terminal or "one shot" evaluation efforts.

Unit One

Module Two

Review

1. Define the following terms:

   A. Mission Statement

   B. Program Goal

   C. Program Objectives

   D. Evaluation Criteria

2. Tests are often used as measures of program outcome. Another way to measure outcome is by ____________ outcomes.

3. Assume you are asked to submit a grant proposal for a program to assess the vocational potential of severely disabled persons. Further assume that the program will be for two years and that you have unlimited resources.

   Write a program mission statement

   Write a program goal statement

   Write a program objective

4. What criterion will you use to measure attainment of your program objective?

References

Bennett, E., & Weisinger, M. **Program Evaluation: A Resource Handbook for Vocational Rehabilitation**. New York: ICD Rehabilitation and Research Center, 1974.

Unit One

Module Three

Program Evaluation in a Political Context

The major purpose of any program evaluation is to make a judgment regarding the program being evaluated. Suchman (1967) has stated that "unlike the basic researcher the applied researcher must be constantly aware of the potential utility of his findings. Only rarely can he take consolation in the fact that the operation was a success but the patient died" (p. 21). Suchman clearly admonishes the evaluator to take responsibility for assuring that the evaluation activity will have applicability. The evaluator is not, however, allowed the liberty and control that the basic researcher takes for granted. Rather the program evaluator is forced to function within established organizational boundaries.

Gurel (1975) has stated that every organization has two opposing superordinate goals, stability/survival and growth/change. Evaluators tend to identify with change, administrators with stability. Overt administrative goals may be stated in terms of growth and change, whereas the more important goals, from an administrative point of view, are stability and survival. Unless an organization has a history of continuing program evaluation activity and consequent organizational modification, an evaluation is likely to be called for only because of outside (i.e., funding source) pressure. Any program evaluation which serves a pure accountability (i.e., survival) function often focuses on the wrong questions, is "quick and dirty" and has no program improvement implications. Roessler has refined the "political" issue by pointing out that when a program evaluator is faced with program goals stated in political terms it is likely that there will be "multiple interpretations of ambiguous objectives, lack of clearly defined social change criteria, conflicting expectations of evaluation by management, and problems of the evaluation meeting its social responsibility while at the same time meeting its objectivity" (Roessler, 1974, p. 35). Weiss (1973) has alluded to the difficulties in conducting a program evaluation when the evaluation is mandated by the funding source or a higher order authority. She indicates there often is difficulty in transferring legislated evaluation mandates into practice because the legislation either doesn't clearly state the operations to be used or states the operations in impractical ways. The difficulty that state vocational rehabilitation agencies have had in addressing federally legislated program evaluation "standards" is a case in point.

That evaluations are often used to maintain and institutionalize existing programs is well known. An example where an evaluation activity failed to impact policy is in the failure of the Nixon administration to incorporate the findings and recommendations of its own blue ribbon commission on pornography. In this case the evaluation was simply discounted and the status quo maintained.

While there will likely always be differences in implicit and explicit agency goals (indeed that is the one reason an evaluation has to make compromises in scientific rigor), there are steps an evaluator can take to insure evaluation utility. First, the evaluator needs to recognize that different value systems are operating in any evaluation scheme. The evaluator must consider the values of the institution under evaluation, the sponsoring authority, the subjects of the evaluation, and the evaluator's own values. All impinge upon the evaluation activity. Second, the evaluator can function as a consultant to the agency in the early stages of the evaluation activity. Indeed, Baptista (1972) remarked that the "program evaluator will find himself acting not only as an evaluator, but also as a participant in program planning and even policy formulation" (p. 9). Third, the evaluator should be aware that while evaluation often functions to suggest agency improvement, it also is often used to justify how resources are to be redistributed. Such redistributions are inherently tied to moral-ethical considerations which transcend scientific objectivity. Fourth, program evaluation, in that it provides information for decision making, is a political process. Often evaluators, especially those utilizing a research approach, are more attuned to the nuances of research than to political reality. Evaluators need to be aware that program administrators have other legitimate concerns, for example, maintenance of the organization, in addition to the hope that the program meets its service goals.

Finally, in order to enhance the impact of program evaluation on program policy Scriven (1972) has suggested that before attempting to analyze the degree to which the program has achieved its goals the evaluator should first attend to how good the goals are. It is not inconceivable for program goals to be tied to political values which may preclude any objective evaluation. If the evaluation is to have more than a cursory effect on policy it is necessary that program administrators have a clear stake in the evaluation activity.

Sources of Resistance to Program Evaluation

It should be readily apparent that program evaluation serves several functions. While the most obvious function is the improvement of services to program recipients, program evaluation is also used as a management tool to monitor/evaluate program staff and to account for the functioning of program managers themselves.

The best kind of evaluation activity, from the standpoint of participant acceptability, is that which addresses clearly defined agency service functions. The closer an evaluation comes to providing an accountability function the more threat those under evaluation feel. In reality the very term evaluation connotes threat. Too often, as Caro (1969) has suggested, evaluation is established for covert reasons (e.g., to settle internal organizational disputes, postpone needed decisions, or place responsibility for unpopular decisions outside the agency) rather than for generally acceptable overt reasons (e.g., feedback to administration for program improvement and dissemination of information to the general public).

If program participants perceive that an evaluation is to be used for covert reasons it is understandable that there will be considerable resistance to the activity. Such resistance is manifested in several ways. Lobenthal (1974) has discussed some not uncommon attitudes program staff have toward evaluation research including: "close ranks, let's not wash our dirty linen in public," "these guys don't understand our problems — and they don't want to," and "the only contribution research can make is showing we're doing something important enough to get more budget support" (Lobenthal, 1974, p. 30). Zusman and Bissonette (1973) have gone as far as to say that resistance to evaluation "will be continuous, infinitely resourceful and ultimately defeating" (p. 113). They have pointed out that such resistance often takes the form of "professionalism" whereby staff stress that the information required by evaluators is confidential. Or there may be administrative resistance, in the form of appeals that the evaluation will cause "work disruptions" or that there is a lack of funds or resources for evaluation activities.

Borgatta (1966) has suggested reasons program personnel give in discounting evaluation findings. He pointed out that programs and organizations often have irrational norms. Convention and authority often function to keep the status quo in spite of evidence that supports change. Among the arguments that program-related staff use to discount evaluation findings suggestive of program change are:

1. program effects are long ranged, they can't be ascertained in six months, one year, two years, etc; or
2. program effects are too diffuse and general to be measured by any specific measurement device; or
3. no instrument or measurement device is sensitive enough to measure small but important program effects; or
4. the measurement process in and of itself disturbs the program processes involved.

Borgatta (1966) has suggested that the above arguments are used **before** the evaluation takes place and stacks the deck against evaluation findings which might suggest program change. Borgatta also suggested rationalization that might be used to discount "negative" evaluation findings **after** the fact; namely:

1. **Some** of the persons exposed to the program made great improvement;
2. some of the persons who most needed the program and would show the most benefits were in the control group; and

3. no difference on outcomes between the program and control group implies need for a more intensive program.

These and other arguments are used when a program evaluation fails to demonstrate the desired program effect or suggests some change which affects entrenched policy. Arguments which counter unpopular findings and consequently serve to maintain the status quo are not uncommon.

Practitioners and Resistance to Program Evaluation

Program evaluation must deal with a mix of professional subgroups, the superordinate authority (program sponsors), program managers (administrators) and program staff (service providers). Each group has its own reasons for evolving or dissolving an evaluation activity. Although a program evaluation is done in the name of program clients and those clients are often asked to complete time consuming evaluation forms, it is not the program clients that are being judged. It is, of course, agency effectiveness and efficiency that is on the line in the evaluation activity. Perhaps the most resistance to a program evaluation activity can be expected from program staff, the practitioners who deliver or provide the program services.

Caro (1969) has pointed out that often administrators will see the value of program evaluation; it is done, after all, to help them reach decisions, whereas practitioners will not. Administrators are generally comfortable with the analytical techniques valued by evaluators; practitioners are not. The issue is basic to philosophical differences between service providers and evaluators and stems from two approaches on how best to understand and help human beings. Practitioners in the helping professions generally ascribe to the idiographic (literally one's own hand) approach, believing that knowledge can best be gained through studies of individuals. This point of view is representative of a humanistic-holistic philosophy which sees the greatest good in the values of the individual. On the other hand program evaluation is steeped in the analytical-mechanistic tradition which may recognize that the individual can only be completely understood as a whole, but believes that the best way an individual can be studied is through analysis of separate functions. Program evaluation is based on the premise that the greatest good is determined through the group. To quote House (1976, p. 75) program evaluation is "based on utilitarian ethics which stipulate that a society is just when its institutions are arranged so as to achieve the greatest net balance of satisfaction summed over all individuals." There is, of course, a great deal of philosophical overlap between service providers and program evaluators. Nevertheless, the basic differences are illuminated when practitioners faced with negative evaluation findings say "if the program improved the life of a single person, it was worth it."

Other differences between program practitioners and evaluation personnel stem from job role perceptions. For example, practitioners have fixed schedules, evaluators do not. It can be disconcerting for the practitioner carrying a large case load to observe an evaluator "thinking." Practitioners have a deep commitment to service, and it is equally disconcerting to have to "deny" services to a very real individual so that the individual can be assigned to a control group for evaluation purposes. Also the practitioner has nothing to **immediately** gain from evaluation research (the best one can hope is that the evaluation will validate one's contribution) and where practitioners value effort and immediate solutions, evaluators often take the longer view and value enduring solutions. Therefore, it is not difficult to understand why service staff are often suspicious of program evaluation. Argyris (1958) has suggested that the practitioner's suspicion of evaluation can function to form an ambivalence toward the evaluator which places the evaluator in a marginal role. It is recognized that the evaluator is a helper, but no one is quite sure how the evaluator is supposed to help. Role marginality can place the evaluator in an odd-man-out position which in turn can limit the evaluator's effectiveness.

Ways to Counter Resistance to Evaluation

There are several things an evaluator can do to counter resistance on the part of program staff. Weiss (1970) has proposed that the evaluator first secure support from the highest levels of administration. This is a crucial step if the evaluation is to have any effect on policy. In gaining administrative support Lacks and Plax (1972) have suggested that evaluators avoid an ivory tower approach and present themselves as interested in agency problems rather than just rigorous research. Second, the evaluator should offer a series of proposals and remain willing to compromise the proposals in consideration of agency goals and resources available. Third, the program staff should be involved, from the start, in the evaluation activity. There are, in turn, several steps to follow in assuring program staff involvement. Meet with small groups of service personnel to describe the need for the evaluation and the necessity for following the research design; encourage suggestions on how the evaluation may best be done (listen carefully because there are sure to be good suggestions); remain neutral as to administration-staff conflicts; and point out the benefits the evaluation will have for the program staff (e.g., improvement in services, documentation of effort, increased professional growth, etc.). In working with practitioners Caro (1969) has recommended that the evaluator present an understanding of the practitioners' situation, a sense of technical competence and personal integrity, and also suggested that evaluation findings be given in small doses, to all participants. Finally, evaluation-based recommendations should include alternatives and the evaluator should be available to help in implementing any needed changes.

Glaser and Taylor (1973) have added empirical support to the above suggestions. They analyzed various applied research studies and were able to differentiate successful from unsuccessful projects. At each stage of the research process, the successful projects had certain qualitative differences.

Idea stage. Successful projects were receptive to ideas generated by administrators and practitioners who were consulted early and continually.

Design stage. The more successful projects exhibited a high degree of conflict between participants at this stage. Glaser and Taylor interpreted that this conflict was issue laden and eventually led to compromise and the development of coping abilities among participants.

Funding stage. More successful projects had the host institution contributing to the research effort.

Research stage. The more successful projects had a researcher, usually fulltime, who was committed to maintaining two-way communication. These researchers also made use of program staff suggestions while the research was under way. Glaser and Taylor noted that researchers in both successful and unsuccessful projects believed they were subject to program staff sabotage.

Dissemination of findings stage. Successful projects published monographs, journal articles, and carried out workshops. According to Glaser and Taylor the journal articles were seen by administrators as being only moderately effective. Direct two-way communication was preferred.

Utilization stage. In the final analysis it was the practitioners who determined if the research findings would be adopted. The fact that practitioners had been involved in the research affected utilization of the findings.

Glaser and Taylor's (1973) findings clearly suggest that staff resistance to evaluation activities is likely to take place. However, if the evaluator recognizes that such resistance will take place and understands the reasons behind the resistance, the evaluator is in a better position to take appropriate steps to minimize evaluation sabotage. According to Glaser and Taylor (1973) the key to successful program evaluation is continual interaction between evaluators and all levels of program staff.

Unit One

Module Three

Review

1. According to Gurel, organizations have two opposing superordinate goals, namely,

1. and

2.

2. What are four steps evaluators can take to insure evaluation utility?

1.

2.

3.

4.

3. According to Borgatta, what are some reasons program personnel might give in discounting evaluation outcomes?

1.

2.

3.

4. Give some ways to counter resistance to evaluation.

1.

2.

3.

References

Argyris, C. Creating effective research relationships in organizations. **Human Organization**, 1958, **17**, 34-48.

Baptista, J. (Chm). **Program evaluation: A beginning statement**. Stillwater, Ok.: Clearing House on Rehabilitation Materials, 1972.

Borgatta, E. Research problems in evaluation of health service demonstrations. **Millbank Memorial Fund Quarterly**. 1966, **44**, 182-201.

Caro, F. Approaches to evaluative research: A review, **Human Organization**, 1969, **28**, 87-89.

Glaser, E., & Taylor, S. Factors influencing the success of applied research. **American Psychologist**, 1973, **28**, 140-146.

Gurel, L. The human side of evaluating human services: Problems and prospects. In Guttentag, M., & Struening, E. (Eds.), **Handbook of Evaluation Research** (Vol. 2). Beverly Hills, Calif.: Sage Publications, 1975, 11-28.

House, E. Justice in evaluation. In Glass, G., (Ed.), **Evaluation Studies Review Annual**. Beverly Hills, Calif.: Sage Publications, 1976, 75-100.

Lacks, P., & Plax, K. **A research program for a vocational rehabilitation agency**. St. Louis, Mo.: Jewish Employment and Vocational Service, 1972.

Lobenthal J. Designing research in corrections: An abbreviated tour guide. **Federal Probation**, 1974, **38**, 29-36.

Roessler, R. Issues in initiating large scale rehabilitation research. In Rubin, S., (Ed.), Issues in program evaluation research in state rehabilitation agencies. **Conference Proceedings**. Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1974.

Scriven, M. The methodology of evaluation. In Weiss, C., (Ed.), **Evaluating Action Programs**. Boston, Mass.: Allyn and Bacon, 1972, 123-136.

Suchman, E. **Evaluative Research**. New York: Russell Sage Foundation, 1967.

Weiss, C. Between the cup and lip. **Evaluation**, 1973, **1**, 49-55.

Weiss, C. The politicization of evaluation research. **Journal of Social Issues**, 1970, **26**, 57-68.

Zusman, J., & Bissonette, R. The case against evaluation. **International Journal of Mental Health**, 1973, **2**, 111-125.

Unit One

Module Four

Program Evaluation and the Scientific Method

> There are many ways to evaluate a program, including judgments on the part of program management, satisfaction by the beneficiaries or consumers of the program and the benefits attributed to it where such benefit-cost analyses range from an anecdotal approach to a highly sophisticated amalgamation of econometrics, experimentation and evaluation research (Perloff, Perloff, & Sussna, 1976, p. 570).

Assuming organizational support for program evaluation and that the program evaluator has been able to deduce a program's mission, goals and objectives, the evaluator is faced with choosing an evaluation strategy. The tactics, or evaluation techniques will follow from the overall strategy. No matter how it is construed program evaluation is faced with answering questions, and for some time the most accepted strategy for answering questions in a factual manner has been the scientific method.

The Scientific Method

Contrary to popular opinion the scientific method has nothing to do with people in white coats, esoteric formulas or test tubes. Rather, it is a highly developed form of inquiry. The goal of the scientific method is to establish objective facts on which two or more persons can agree. In that program evaluation is tied to judging a program's worth, such a form of inquiry has obvious benefits.

The scientific method is based on a systematic analysis of factual data. Facts are established through highly objective observations (empirical analysis) and experimentation. According to the scientific method a causal explanation is accepted only if it is obtained empirically, can be expressed in quantifiable terms, is capable of being refuted, is the least complicated explanation, and the method establishing the explanation is open to public inspection. Use of the scientific method has greatly added to our knowledge base. It should be noted however, that the scientific method is not the only way to establish knowledge and that some things (e.g., religious explanations) may not be amenable to analysis via this method.

Program Evaluation and Research

Research and the scientific method are often considered synonomous whereas evaluation connotes personal values and judgments of worth. **Evaluation research** refers to scientifically establishing facts so that value judgments can be made. Within the discipline of program evaluation there is an effort to distinguish between program evaluation per se, evaluation research, and traditional research. We see program evaluation as an overall term encompassing many methods of judging program worth, evaluation research as the preferred strategy and traditional research providing the tactics for evaluation research.

> **Evaluation research refers to scientifically establishing facts in order to make value judgments about programs.**

The main difference between evaluation research and traditional scientific research is that traditional research may have as its goal the discovery of new knowledge, irrespective of whether or not that knowledge is immediately applicable to some problem. This inquiry, called basic research, is seldom conducted in the arena of program evaluation. Rather, evaluation researchers seek to apply established techniques and test different service methods (i.e., programs) in real life settings.

The Tactics of Evaluation Research

Experimental research. According to McGuigan (1968), the "most powerful application of the scientific method is experimentation" (p. 3). The way the experimental method is applied and the reasons it is so powerful follow. First, a tentative solution to a problem is suggested. This tentative solution is called a hypothesis. In program evaluation, a hypothesis might be "persons who undertake vocational training under program A, will upon program completion be more employable than persons in program B." Second, persons are assigned to either program A or B. The way persons are assigned to the groups is important. By using random assignment (each person having an equal chance of being selected for either group), the ultimate experimental findings will be more powerful. People in group A receive the new program or treatment. The treatment (e.g., group counseling, increased services, etc.) is referred to as the **independent** variable. The measure which will determine the degree to which a person is "employable" is chosen. This measure, the **dependent** variable, is quantitative in nature; the numbers obtained relate to the degree any one person is employable. Assuming random assignment and everything else being equal, the only difference between program A and B is the experimental treatment (i.e., the "program"). Hence, if it can be shown that after the program is completed, persons in program A (experimental group) are indeed more employable than persons in program B (the control group), we can say that program A, relative to program B, resulted in increased employability. Depending on certain conditions (e.g., how persons in both programs were selected), we may also be able to say that program A was the cause of the intended effect, increased employability. Establishing cause and effect relationships is the usual purpose of experimentation.

Correlational Research. Research which focuses on associations, or relationships between two or more variables is correlational in nature. Correlational methods were developed by scientists who had an interest in individual differences. Contrary to experimentalists who hope to discover universal laws, correlationalists are interested in discovering the interrelationships between environment, program and person characteristics. The relationship between social status and psychological functioning is a question addressed through correlational methods. Cronbach (1957) has suggested that the correlationist assumes a population of persons rich in individual differences and searches for a fixed treatment to best fit the population under study. On the other hand, the experimentalist assumes a fixed population and searches among many treatments to identify that treatment which will best fit the most people.

Correlational research forms the basis of much modern day psychometrics and is exemplified in multivariate research which includes the organization of complex interrelationships. Contrary to most experimental research, correlational research does not imply causation. Correlational research is just as scientific as experimental research. In fact, Cronbach (1957) has pointed out that correlational methods can be used to test predictions that some intervention will result in a specified change. In evaluation research correlational techniques can be advantageous when (1) experimental manipulation is impossible because of ethical or cost factors, (2) as a first test to determine if a new intervention approach is worth the effort, or (3) when the evaluator's primary purpose is to describe underlying dimensions of programmatic effect.

Case Studies. A third type of research activity is the case study. The case study is an important tool in the helping professions. In fact, the clinical psychological and vocational rehabilitation literature abounds with case study findings. The method probably gains most of its prestige from its use in the medical sciences. In fact one of the more elegant medical models of psychological adjustment, Freud's psychology, was developed almost exclusively using the case study approach.

In program evaluation, the case study has its proponents and is sometimes found in the form of program participant testimonials. In vocational rehabilitation the case study is alive and well as the case review. The main difficulty with case studies, and the reason they are considered "soft studies" by some "rigorous" researchers, is that they are susceptible to subjective interpretations and often lack generalizability. Also data generated through case studies often resist translation from qualitative form into a more manageable quantitative form. However, with the development of such techniques as the multiple baseline, case studies can make valuable contributions to evaluation research.

> **There are three broad tactical approaches to research-evaluation: experimentation, correlational analysis and case studies.**

Advantages and Disadvantages of the Research Approach to Program Evaluation

The most rigorous type of evaluation research, in terms of drawing causal inferences, is experimental in nature. The advantage of the experimental approach, is that the evaluator can objectively answer the question "Did this program or project really make a difference?"

Evaluation research can take many forms from the classical experimental approach to the analysis of program component interrelationships. However, the tactics of traditional research, and of experimentation in particular, have been criticized by applied program evaluators. They see the experimental approach as too narrow and more suited to the laboratory than to the evaluation of complex social programs. Research as program evaluation is often seen as not viable for the following reasons.

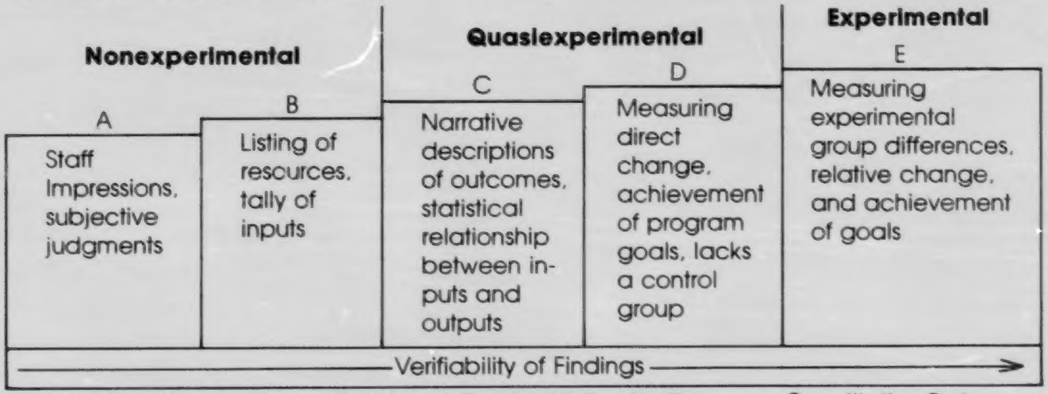
1. Research follows theory, and the relevant questions are asked by the scientist. Program evaluators have to use program goals, which may not be testable. Furthermore, the overall purpose of program evaluation is not to discover new knowledge but to provide information for administrative decisions.
2. Program evaluation is inherently linked to administrative values; an objective/scientific evaluation of a program is not really possible.
3. The experimental method of utilizing a no-program control group is unethical. One cannot withhold services to those eligible and deserving program services.
4. Even if comparison of alternative programs is possible, random assignment of persons to programs is not.
5. Traditional experimental methods seek to control all extraneous variables in assessing broad service programs. One cannot assume that humans are interchangeable and that the treatment will outweigh environmental effects.
6. Research demands that program goals be predetermined and remain the same for the life of the program. In reality goals change as program priorities change.
7. A test of statistical significance is not the same as a test of programmatic significance. Because programs serve large numbers of persons, program effects can often be shown to be statistically significant yet have little practical meaning.
8. Reliance on experimental rigor causes the evaluator to disregard rich subjective and impressionistic data.
9. A research evaluation is summative (terminal) in nature. Nothing can be done to improve the program or help the decision maker while the program is in progress.
10. The program is often forced to fit the evaluation design rather than vice versa.

Contrasting the above points which illustrate the difficulty of conducting research oriented evaluations are some of the advantages of using a research approach to program evaluation.

1. Suchman (1967) has stated that any program intervention implies a causal relationship between the program and some expected outcome. A research orientation is needed to document that relationship.

2. Because the research orientation focuses on causal relationships between program and outcome variables, program failures are as important as program successes. Rather than focusing on **how** a program functions, the research orientation can suggest **why** it worked. The implications for improved programs are obvious.
3. Use of a control group permits comparisons as to what effect the program actually has. The use of nontreatment control groups is highly ethical. The point of the evaluation is to determine if the program is of benefit or at least not harmful. As Boruch (1976) has pointed out, it is unethical to obtain equivocal data through "similar" group comparisons which often belie **any** interpretations. When treatment control groups are not feasible for political or cost reasons; use of other procedures such as alternative treatment groups can maintain the experimental flavor of the evaluation.
4. Randomization of persons is necessary to rule out biases which may be affecting a program's outcome. Because program evaluation seeks to determine the effects of a program, it is important to note whether there are systematic differences between persons eligible for the program and persons selected into the program. When resources are scarce and the population to be served large and equally deserving, random assignment of persons to the program is the most democratic of procedures.
5. The individual is unique and extremely complex. Only by considering the aggregate do human actions become clarified. For better or worse, programs are large scale interventions dealing with large groups of persons. Research methodology, based on probability theory, remains the most efficient and economical way to measure a program's impact on a population of people.
6. Use of a research approach to program evaluation forces the evaluator and the program staff to focus on measureable program goals.
7. A program evaluation activity usually implies that some change among program participants takes place. In order to measure participant gain (or loss), measurement schemes need to be employed. In order to give meaning to the observed participant change (i.e., amount of change attributed to the program versus change expected by chance), a research approach should be utilized.
8. If the program evaluation findings are to generalize to other similar programs or to the larger program as a whole, then good sampling methods and the experimental method is necessary.

We believe that the scientific method is the best strategy the evaluator can follow. We also view the experimental method as the most desirable tactic to use. There are, however, many instances where the experimental method, in its most pure sense, is not applicable. We therefore feel that the scientific method can best be applied to program evaluation concerns through evaluation research. Evaluation research may be defined as "a variety of empirically oriented research technologies, spanning the gamut from impressionistic procedures, field or survey methods and an armamentarium of sophisticated research and quantitative procedures" (Perloff, Perloff, & Sussna, 1976, p. 570).


Nonexperimental
Quasiexperimental
Experimental
A
Staff Impressions, subjective judgments
B
Listing of rescurces, tally of inputs
C
Narrative descriptions of outcomes, statistical relationship between inputs and outputs
D
Measuring direct change, achievement of program goals, lacks a control group
E
Measuring experimental group differences, relative change, and achievement of goals
Verifiability of Findings
Qualitative Data
Quantitative Data

Tactics of Research Evaluation
Figure 1

In that evaluation research compromises the experimental method, is decision oriented, and emphasizes immediate applicability of results, it differs from traditional research. It is important to note that in this context evaluation research exists on a continuum of objectives, generalizability and verifiability. Figure 1 illustrates this continuum. The most objective, generalizable and verifiable information, other things being equal, is obtained from experimental evaluations. However, in that the scientific method is followed, all of the steps on the continuum are considered evaluation research. In fact, the most comprehensive program evaluation would utilize information from each step in the continuum.

Unit One

Module Four

Review

1. The scientific method is a form of ______________________________

based on a systematic analysis of ______________________________

2. Which of the following is needed before casual inference can be drawn?

   1. Facts are established empirically.

   2. Program outcome is dependent on the director's judgment.

   3. Results are open to public inspection.

   4. The importance of the findings depends on the faith one has in them.

3. Three drawbacks to using a traditional research approach in program evaluation are:

   1.

   2.

   3.

4. Three advantages to using a traditional research approach in program evaluation are:

1.

2.

3.

5. Evaluation research may be defined as ____

____

6. What are three tactical approaches to evaluation research?

1.

2.

3.

References

Boruch, R. On common contentions about nonrandomized field experiments. In Glass, F., (Ed.), **Evaluation studies review annual.** Beverly Hills, Calif.: Sage Publications, 1976, 158-194.

Cronback, L. The two disciplines of scientific psychology. **American Psychologist**, 1957, **12**, 671-684.

McGuigan, F. **Experimental psychology [2nd ed.]**. Englewood Cliffs, N.J.: Prentice Hall, 1968.

Perloff, R., Perloff, E., & Sussna, E. Program evaluation. In Rosenzweig, M., & Porter, L., (Eds.), **Annual review of psychology [Vol. 27]**. Palo Alto, Calif.: Annual Reviews Inc., 1976.

Suchman, E. **Evaluative research**. New York: Russell Sage Foundation, 1967.

Unit II

Concepts in Sampling

Table of Contents

Module

Introduction to Sampling 35

Objective: To present advantages of the sampling method for program evaluators

Statistical Signs and Symbols 39

Objective: To provide definitions for basic notation used in evaluative research

Some Sampling Concepts 43

Objective: To provide definition and experience in the computation of quantities useful in sampling

Selecting the Random Sample 47

Objective: To stimulate thinking regarding various techniques of randomization

Variability in Random Samples 51

Objective: To emphasize the difference between unit-to-unit variability and sample-to-sample variability

Components of a Sampling Investigation 53

Objective: To define and stimulate thinking regarding various components of the sampling process

Estimation of Sample Size 55

Objective: To provide the information and skills required for the selection of sample size in a study to estimate a proportion

Sample Size for Estimating a Mean 61

Objective: To provide the information and skills required for the selection of sample size in a study to estimate a mean

The Pilot Study 67

Objective: To emphasize the utility of a pilot study in survey sampling

Sample Size with Several Subpopulations or Several Items 69

Objective: To stimulate thinking regarding the feasibility of stringent precision requirements for multi-item or multi-population surveys

Unit II

Module 1

Introduction to Sampling

In the early 1800's, the results of national elections in the United States were not available to some areas of the country for many weeks after the votes were actually counted. Today it is possible to accurately predict the outcome of an election even before all ballots have been cast. In addition to the common practice of declaring a winner before all votes are counted, media network sampling experts and political pollsters continuously bombard the public with the results of various opinion polls telling us what the majority of voters think about certain issues or which candidate is in the lead for some party's nomination. Over the past several elections these polls have been highly accurate in predicting the actual vote. This feat is astounding when one considers that the poll takers are able to predict, within one or two percentage points, how millions of voters will cast their votes by surveying only a tiny percentage of the people who might vote. The reason modern opinion polls are so successful is primarily because of increased knowledge in **survey sampling** techniques. Knowledge of the basics of sampling can also be very important in program evaluation.

Obviously it is impractical to ask every person in the United States for whom he or she might vote; likewise in a state rehabilitation agency, it is economically infeasible to conduct a detailed review of every case served. In practice, both opinion surveyors and program evaluators often take a **sample** of the total population. It is, of course, very important that the sample be representative of the total population about which inferences are to be made. This is the basic problem of sampling and the techniques presented in this unit help to insure an adequate and representative sample.

Cochran (1963) lists four advantages of sampling over total enumeration. As applied to program evaluation in vocational rehabilitation they might be paraphrased: (1) reduced resource requirements; (2) more timely results; (3) potential for greater scope; and (4) greater accuracy of results.

Reduced resource requirements. It is clear that less man power and money are required to review a sample of (say) ten percent of a counselor's cases than would be required to review the whole caseload. It should also be emplasized that this reduction in cost can often occur concurrently with negligible loss in retrieved information.

More timely results. In state vocational rehabilitation agencies, data processing is often a time consuming process. Further, much evaluation data must be analyzed relatively quickly or its usefulness to management personnel diminishes greatly. By using sampling techniques rather than total enumeration the resulting data may be analyzed more quickly so that the results are available to management in time for utilization in planning and decision making.

Potential for greater scope. Some program evaluation studies are precluded due to their requirements for highly trained personnel, specialized equipment, or facilities with limited availability. Through careful use of sampling technology such studies may be conducted and their results generalized to larger populations. Thus, knowledge of sampling technology increases the program evaluator's flexibility regarding the types of information which are feasible to collect.

Greater accuracy of results. Because of reduced time and cost requirements and the potential for collection of more specific information, results of a sampling review can be more accurate and meaningful than results of a total caseload review. Further, for special studies, personnel of higher quality can be hired and supervised more closely than would be economically feasible in a total enumeration study. Thus results based on detailed information from a sample may be more accurate than results based on more general information from the total population.

In conducting a sampling study, it is first necessary to define the total population under consideration. A listing of this population is referred to as the sampling **frame**. For example, an evaluator may define the population to be the collection of all persons served by the state rehabilitation agency during 1976. The evaluator then determines the proper size of the sample to be drawn (see Module V). Suppose the total population is 10,000 persons and the evaluator decides a sample size of 500 persons is adequate. Actually selecting the sample involves some thought. Merely taking the first 500 client cases off the agency's master list may produce a sample which is not representative of the agency's total caseload. Perhaps the first 500 clients were different from the total caseload in some respect. For example, suppose most of these 500 were suffering from alcohol abuse and had applied for service immediately following the New Year's Holidays. In order to assure that the sample is representative of the total population, an evaluator must take a **random** sample. In sampling theory "random" often means that every person in the total population has an equal chance of being included in the sample. Alternatively, it may be said that a random sample of size n is such that every possible sample of size n has an equal probability of being selected. Techniques for insuring the randomness of a sample are discussed in this unit.

In sampling from a population we are, of course, interested in making **inferences** based on the sample about the total population. In following sections, statistical methods for making such inferences are introduced. A **statistic** is an estimate made from a sample which corresponds to some population value or parameter. Thus a sample mean (a statistic) denoted by $\bar{X}$ is an estimate of the population mean denoted by μ (mu). The idea in statistical inference is to determine the amount of error attached to the sample estimate. That is, how much do we expect the sample estimate ($\bar{X}$) to differ from the true population value (μ)? Good sampling procedures help to insure that this error is held to a minimum.

References

Cochran, W.G. **Sampling Techniques**. (second edition) New York: John Wiley & Sons, Inc., 1963.

Unit II

Module 2

Statistical Signs and Symbols

Although the program evaluation specialist has no need to become an expert on mathematical and statistical notation, it is often convenient to have familiarity with selected mathematical symbols. Basically, this familiarity can serve three useful purposes:

1. Some knowledge of mathematical symbols can facilitate individual study regarding new data analysis techniques or reading technical reports of data analyses from other agencies.
2. Program evaluators may find it convenient, as do statisticians, to write results of their data analysis using some mathematical symbols.
3. Familiarity with a few mathematical symbols facilitates the communication of program evaluators with data analysis experts.

The following short collection of symbols represents those most commonly used in data analysis and evaluation research.

x_i The subscripted variable represents the score of the i^th^ subject on a given measurement instrument. Consider, for example, the IQ scores of five individuals given in the table below:

Table 1

Subscripted Variables

Subject	Score Symbol	Score value
1	x_1	115
2	x_2	120
3	x_3	117
4	x_4	130
5	x_5	144

In this example, the variable x represents IQ. Actually, it could be height, weight, time in referral status, etc. The point is that x_1 represents the score of subject 1, x_2 represents the score of subject 2, etc. In general, x_i represents the score of subject i. For example, in the table above, the IQ of subject 3 is represented by x_3 and has an observed value of 117.

Σ The Greek capital letter sigma (Σ) is used to denote summation. It is often seen written in the form $\sum\limits_{i=1}^{n} x_i$. The small symbols above and below the Σ are the **limits of summation** and indicate the numbers to be summed. $\sum\limits_{i=1}^{n} x_i$ is read "sum of x_i from i = 1 to n" and is defined by the following equality:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$$

For the above data,

$$\sum_{i=1}^{5} x_i = 115 + 120 + 117 + 130 + 144 = 626.$$

Also $\sum_{i=1}^{3} x_i = 115 + 120 + 117 = 352$.

Sometimes, when there is no confusion as to what the limits of summation are to be, they and the subscripts on the variable are omitted so that $\Sigma x = \sum_{i=1}^{n} x_i$.

x_i^2	The superscript ^{2} indicates that the quantity represented by x_i is to be squared, that is, multiplied by itself. Note: $x_i^2 = x_i$ times x_i For example, for the data above $x_2^2 = 120 \times 120 = 14400$
$\sqrt{}$	The square root symbol means take the square root of the quantity under the symbol. The square root of a quantity x, denoted $\sqrt{x}$, is a number such that $\sqrt{x} \times \sqrt{x} = x$. For example, from the data in the table, the square root of x_5 is given by $\sqrt{x_5} = \sqrt{144} = 12$ since $12 \times 12 = 144$. Note that "$\sqrt{}$" is the inverse operation of "^{2}" since $\sqrt{x^2} = x$.
( )	An algebraic operation in parentheses is performed before operations outside the parentheses. For example $(4 + 6)^2$ means first add 4 and 6 then square the resulting number of $(4 + 6) \times (4 + 6) = 10 \times 10 = 100$. Note also that $\sqrt{(9 + 7)} = \sqrt{16} = 4$.
$\bar{x}$	The symbol $\bar{x}$, read "x bar," denotes the quantity $\frac{\sum_{i=1}^{n} x_i}{n}$ which is the average of n observations. This quantity is often referred to as a **sample mean**.
n!	The symbol n!, read "n factorial,"denotes the quanitity n x (n-1) × (n-2) × ... × 1. For example, $4! = 4 \times 3 \times 2 \times 1 = 24$. n! may be interpreted as the number of possible arrangements of n objects. There are 3! = 6 distinct arrangements of 3 objects.

Review

The following table gives the lengths of time (in weeks) that 5 rehabilitation clients spent in status 00.

Client	Symbol	Time
1	x_1	4
2	x_2	6
3	x_3	2
4	x_4	1
5	x_5	3
	$\sum_{i=1}^{n} x_i$ =	16

1. Compute the following for the data in the table above.

a. $\sum_{i=1}^{n} x_i$ Answer: 16

b. $\sum_{i=1}^{5} x_i^2$ Answer: 66

c. $\overline{x}$ Answer: 3.2

d. $\frac{\Sigma (x_i - \overline{x})^2}{4}$ Answer: 3.7

2. For any constant number c, the following facts are true.

$$c \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (cx_i)$$

and

$$\sum_{i=1}^{n} c = nc$$

a. Find $\sum_{i=1}^{5} (3x_i)$ Answer: 48

b. Find $\sum_{i=1}^{5} 2$ Answer: 10

3. Compute 3! and list the 3! arrangements of the letters A, B, and C. Answer: 3! = 3 • 2 • 1 = 6

ABC, ACB, BAC,
BCA, CBA, CAB

Unit II

Module Three

Some Sampling Concepts

Recall that the word **population** is used to denote the collection of **elements** from which a sample is to be taken. N (capital N) is used to denote the number of elements in the population while n (small n) is used to denote the number of elements in the sample.

In planning a sample survey, certain characteristics of the population elements are determined to be of importance. For example, if we are interested in estimating the mean earnings of rehabilitation clients referred to a state agency, the characteristic of interest might be Weekly Earnings at Referral. The values of this characteristic for the elements in the population are denoted by $Y_1, Y_2, \ldots Y_N$ In this example, $Y_1, Y_2, \ldots Y_N$ denote the weekly earnings at referral of the 1st, 2nd, ... and Nth rehabilitation clients (respectively) in the population. The **population total** for the characteristic of interest is denoted by Y and defined as:

$$Y = \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + \ldots + Y_N.$$

The **population mean** is denoted by $\overline{Y}$ and is defined as:

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{Y_1 + Y_2 + \ldots Y_N}{N}.$$

Similarly, for a sample of size n,$(y_1, Y_2, \ldots Y_n)$ the sample total is $\sum_{i=1}^{n} y_i = y_1 + y_2 + y_n$

and the sample mean is $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n} = \dfrac{y_1 + y_2 + \ldots + y_n}{n}$.

The sample standard deviation, denoted sd, is the most commonly used indicator of the variability or dispersion among the sample scores. The standard deviation is defined by the following expression:

$$sd = \sqrt{\frac{\sum_{i=1}^{n} (y_i-\overline{y})^2}{n-1}}$$

This quantity is often computed using the following five steps:

1. Substract the sample mean, $\overline{y}$, from each score:
   $(y_1-\overline{y}), (y_2-\overline{y}), \ldots (y_n-\overline{y})$.
2. Square each of these differences:
   $(y_1-\overline{y})^2, (y_2-\overline{y})^2, \ldots (y_n-\overline{y})^2$.
3. Sum the squared differences:
   $(y_1-\overline{y})^2 + (y_2-\overline{y})^2 + \ldots. + (y_n-\overline{y})^2 = \sum_{i=1}^{n} (y_i-\overline{y})^2$.

4. Divide this sum by the sample size minus 1, n-1:

$$\frac{\sum_{i=1}^{n} (y_i-\bar{y})^2}{n-1}.$$

This quantity is called the **sample variance**.

5. Take the square root of the sample variance:

$$sd = \sqrt{\frac{\sum_{i=1}^{n} (y_i-\bar{y})^2}{n-1}}.$$

Note: It is traditional that capital letters refer to population characteristics while lower case letters refer to those of a sample.

Simple Random Sampling. The number of different samples of size n which may be drawn from a population of size N is denoted by ${}_NC_n$ and may be computed from the formula:

$${}_NC_n = \frac{N!}{n!\,(N-n)!}$$

,which is read "number of combinations of N things taken n at a time." For example, if the population contains four units, A, B, C and D, there are $\frac{4 \cdot 3 \cdot 2}{2 \cdot 1\,(2 \cdot 1)} = 6$ ways of choosing a sample of size two. The six possiblities are AB, AC, AD, BC, BD, and CD.

Recall that the symbol N!, read "N factorial," is defined by the following equality:

$$N! = N \times (N-1) \times \ldots \times 1.$$

Likewise, $n! = n \times (n-1) \times \ldots \times 1$. Thus it is easy to compute ${}_NC_n$, the number of different samples of size n which can be taken from a population of size N.

Simple random sampling (SRS) is a procedure for selecting a sample of size n from a population of size N so that each possible sample of size n has an equal chance of being selected. Thus to select a simple random sample of size two from the population of four letters (A,B,C,D) we must choose a pair from the six possibilities in a way such that each pair has an equal chance (1/6) of being selected. One method might be to number the **pairs** from 1 to 6 and roll a die to determine our choice. In practice, it is usually more feasible to select the sample one unit at a time. That is, one letter is chosen from the four possibilities followed by the selection of a second letter from the **remaining** three possibilities. This is called sampling **without replacement**.

Review

1. Suppose the average weekly earnings at referral for the population of all clients referred to and accepted for services by a state rehabilitation agency in FY 1976 are \$40. If there were 10,000 such clients, then the population total (i.e., the sum of the weekly earnings for all 10,000 clients)
   Y =
   Answer: \$400,000.

2. Suppose a sample of 10 clients was chosen from this population and their weekly earnings at referral are given by: \$50, \$40, \$30, \$50, \$35, \$65, \$100, \$20, \$30, \$40.

   The sample mean $\overline{y}$ =
   Answer: \$46.

3. If the sample mean is \$46 then a reasonable estimate of the population total would be:
   Answer: \$46 × N = \$460,000.

4. Suppose a supervisor is interested in assessing the satisfaction of a particular counselor's clients. He wants to interview a random sample of twenty clients from the counselor's caseload (N clients). Which of the following procedures would result in a simple random sample of size 20?
   a. The supervisor selects the first twenty cases on the counselor's master list.
   b. The supervisor selects every 20th case on the master list.
   c. The supervisor thoroughly shuffles the cases on the master list then chooses the first twenty cases.
   d. From a table of random numbers (see Unit II, Module Four), the supervisor selects twenty numbers from 1 to N and selects the cases on the master list corresponding to those numbers.
   e. The supervisor tells his secretary to pull 20 of the counselor's current case files.
   Answer: c and d will produce simple random samples of size 20.

5. A district office has eight counselors. How many possible samples of size 3 can be selected from this population? Answer: 56.

Unit II

Module Four

Selecting the Random Sample

From time to time in the course of events, the responsibility for conducting a case review audit falls upon Charlie Numbercruncher, our friendly neighborhood program evaluation specialist. Since Charlie has had all too much experience with this unpleasant task he knows he can feasibly audit only 100 cases in the three days he has allocated for each district office. A recurrent problem has been the selection of cases to audit. Past experience has indicated that allowing the counselors to select the cases tends to give biased results. This year, Charlie plans to select a random sample of 100 active cases from the master list of each district office.

Charlie asked P.J. Swagdata, the complement of the program evaluation team, to investigate the three most common methods for randomly selecting a sample. These three methods are:

(1) a table of random numbers,
(2) a sample randomly selected by a computer,
(3) various home-brew randomization devices.

P.J. made the following notes:

Table of random numbers. A table of random numbers is possibly the most straightforward technique for selecting a random sample of cases from a master list. First, the cases on the list are numbered sequentially. Suppose there were 4,000 active cases on the list in the first district office. Second, a table of random numbers is entered at an arbitrary point and the desired number of 4 digit numbers are read off, deleting any which are over 4000. For example, consider the abbreviated table of random numbers in Table 1.

Starting in column 2, row 11 and reading down we get the following random numbers: **9315**, 1940, **4771**, **8871**, **8595**, **6187**, 3616, 3004, 3298, 3427, 2223, etc. The bold faced numbers are greater than 4000 and are therefore discarded. After 100 numbers have been selected in this manner, the files for cases on the master list corresponding to these numbers are pulled for audit. Clearly the abbreviated table given here is not large enough to give the 100 numbers needed. Larger tables are available as appendices in many statistics textbooks or in books of mathematical tables such as the commonly available CRC Handbook:

Standard Mathematical Tables
Samuel M. Selby, Editor,
The Chemical Rubber Company, Cleveland, Ohio.

Computer Generated Sample. Virtually all modern computing systems have some sort of built-in random number generating procedure. This procedure may be used as the basis of a program to print out a random sample of cases from a master list tape. If an automated data processing system is available, the computer generated sample is usually the most efficient method for selecting cases for audit or review.

Home-brew Randomization Devices. When only a small random sample is required, the inventive investigator can nearly always find ways of randomly selecting a sample. For example, to select a sample of say 10 cases from a sequentially numbered master list containing 4000 names, we may use the following device. Place 4 coffee cups in a row. Put 5 slips of paper in the first numbered 0, 1, 2, 3, and 4 respectively. Place slips numbered from 0 to 9 in each of the remaining cups. Stir each cup with any available utensil. Draw one number from each cup and read the 4 digit number corresponding to a case to be selected for the sample. Replace the slips in their respective cups and continue until the desired sample size is reached. Note that some numbers will have to be discarded, e.g., 4091.

Table I

Random Numbers

76484	83095	41282	48852
51764	23954	15324	81036
62749	33349	67455	55698
70853	00414	73689	94686
69123	85183	78599	85691
29459	20966	53998	19380
60725	10204	69175	30197
86799	39741	40321	63908
24052	88913	70212	81172
07767	97531	07586	97569
14474	93153	26053	25650
81477	19401	70255	36507
58981	47716	23038	27154
80615	88717	33872	41629
57078	85952	61034	26440
25669	61871	46555	60141
23551	36167	12243	49722
43304	30049	79764	21508
21751	32986	83573	88092
58726	34270	40682	67083
27059	22239	82735	83865
32369	72103	77745	81446
26183	59964	26289	41841
00265	68523	10969	76782
19860	72190	34817	43005

As far as randomness of the sample is concerned, all three procedures are equally effective methods of selecting a sample. The appropriate method for use therefore depends on practical considerations such as sample size and the availability of computing facilities, tables, and paper cups.

Review

1. Using the table of random numbers given in this section, find 10 random numbers between 1 and 78 inclusive. Describe your method.

2. Describe a "home-brew" randomization procedure different from the one given which you could use to find 50 random numbers between 1 and 1000 inclusive.

3. Describe a method of selecting a sample of size 10 from a list of 100 names which does **not** result in a random sample. Why is the sample not random?

Unit II

Module Five

Variability in Random Samples

There are two types of variability commonly associated with random samples. The first and most commonly recognized is **unit-to-unit variability** or variability between the units of the sample. This type of variability is usually measured in terms of the sample standard deviation or the sample range. The second type of variability is called **sample-to-sample variability** and refers to the differences between the characteristics of samples drawn from the same population. The local program evaluation team had an experience which illustrates both types of variability.

Charlie and P.J. were visiting a district office with the specific purpose of checking compliance with regulations regarding certification of eligibility. They divided up the caseloads and from each caseload a random sample of 10 cases was selected and checked to see whether or not a certification of eligibility was included in the case record. By some miracle of modern technology one counselor's caseload was reviewed by both P.J. and Charlie. Their samples had only one case in common and their results were correspondingly different. Charlie's sample resulted in 10 cases, all of which contained the certification of eligibility, giving the counselor a rating of 100%. P.J.'s sample of 10 cases, however, contained three cases with no record of certification of eligibility even though the cases were currently receiving rehabilitation services. Thus, P.J.'s rating for the counselor was 70%. The difference between the two ratings is an indication of **sample-to-sample variability**.

The counselor is well aware of the problems in his caseload and knows precisely the proportion of his caseload with no certification of eligibility, namely, 1%. The difference between a sample parameter and the true population parameter is another indication of sample-to-sample variation although it is not usually available. Some **sample-to-sample variability** is always present and in most cases the only way to reduce it is to increase the sample size so that there is a smaller chance of selecting an extremely unrepresentative sample.

The concept of sample-to-sample variability is closely associated with that of a confidence interval. A confidence interval on the proportion of satisfied clients in a rehabilitation agency is an interval defined by two end-points (say .6 and .8) about which a statement similar to the following might be made: "The probability is .95 that the true proportion of satisfied clients in the agency is between .6 and .8." This confidence interval (.6, .8) is computed from observations from a sample. If there is large sample-to-sample variability among sample proportions, then a 95% confidence interval on the true proportion must be very large. Likewise, large confidence intervals indicate large sample-to-sample variation.

Sample-to-sample variation is also closely related to unit-to-unit variation. When sampling for a mean such as the mean time from referral to acceptance, a measure of the unit-to-unit variability is the sample standard deviation, sd. The corresponding measure of the sample-to-sample variation in the sample mean based on a sample of size n is given by the **standard error of the mean**, $sd/\sqrt{n}$. Thus there is a direct relationship between unit-to-unit variation and sample-to-sample variation. This relationship also makes clear the fact that increasing the sample size, n, decreases the sample-to-sample variation of the estimate of the population mean.

Review

1. The sample standard deviation is a measure of ______________________________

____.

2. The standard error of the mean is a measure of ______________________________

____.

3. If the unit to unit variation is high then the sample to sample variation may be lowered by______________ the sample size.

4. Large sample to sample variation of the sample mean indicates a ____________ 95% confidence interval on the population mean.

Unit II

Module Six

Components of a Sampling Investigation

There are six basic components to any investigation involving the use of sampling techniques. It is essential that the evaluator include careful consideration of each of these components at the planning stages of the proposed investigation.

Population. The population for which inferences are to be made must be carefully defined with consideration given to the resources available and the generalizability desired. Suppose that an evaluator wanted to estimate the number of rehabilitation clients with police records. This evaluator must specify exactly what population he is considering. One such population would be all currently active rehabilitation clients in the U.S. which have been accepted for services. Of this population the researcher may want to estimate the number convicted of felony offenses prior to acceptance for rehabilitation services.

Sampling Unit. Prior to the selection of a sample the elements of the population must be partitioned into nonoverlapping sampling units. In most studies, the sampling unit is obvious. For example, in the population of currently active rehabilitation clients, the sampling unit is a client. In an investigation to estimate the number of disabled people in a given town the sampling unit might be a family or a dwelling.

Sampling Frame. The sampling frame is a complete listing of all the sampling units in a population. Construction of the frame is sometimes the most difficult part of a sample survey. Further, the accuracy of the sampling frame often determines the success or failure of the entire sampling effort. To sample from the population of active rehabilitation cases in some states, the sampling frame is ready-made — the agency master list of clients.

Sample. The sample is selected directly from the sampling frame. It is sometimes advantageous to select a stratified sample. That is, select a separate sample from each of several distinct subpopulations (strata) of the total population. These subpopulations may be defined by geographical location or other client demographic characteristics. Whatever the basis for stratification, the sampling frame must contain sufficient information to define the strata. The selection of samples within strata is basically the same as selecting a single random sample. The actual selection of random samples is discussed elsewhere in this unit.

Measurement. In every sample survey, the units in the sample must be measured in some way. Although the concepts of reliability, validity and precision are discussed elsewhere, they are important here also. Without adequate measurement technology the sample survey may be either overly expensive or result in information with no practical value. The precision of the measurement device directly affects the required sample size and the precision of the study.

Inference. The last phase of a sampling investigation is usually the drawing of inferences or the generalization of results from the sample to the population being sampled. That is, characteristics of the sample are attributed to the entire population. It is important to realize that the potential inferences are completely dependent on the adequacy of the sampling plan — the sampling frame, the sample selection, and the measurement technology. No amount of statistical extrapolation can bring forth useful information from data collected in a poorly designed survey. To insure that a proposed survey will produce the best possible information, it is nearly always cost-effective to consult an expert in survey sampling.

Review

1. List six components of a sampling investigation.
   1.
   2.
   3.
   4.
   5.
   6.

2. Consider the problem of estimating the satisfaction of rehabilitation clients with agency services. Suppose the study is limited to a particular metropolitan area.
   1. Define an appropriate population.
   2. What are the sampling units?
   3. How would you construct a sampling frame?
   4. What considerations would be important in designing a measurement device?
   5. How would you select the sample?
   6. What inferences would you like to make?

Unit II

Module Seven

Estimation of Sample Size

Suppose state agency program evaluation specialists want to estimate the proportion of cases in a given district office which are out of compliance with a certain facet of the state regulation or policy. For example, it may be desirable to estimate the proportion of IWRP's which have not been signed by the client. Clearly, it is impractical to review every case from the district office, hence some type of sampling procedure must be implemented. In a sampling situation, it is often helpful to be able to estimate the sample size necessary to insure a specified degree of precision or to estimate the expected cost of a review.

Charlie, a program evaluation specialist, has been given the responsibility of estimating the proportion (P) of IWRP's from a specific district office which were not signed by the client. Being the practical type, Charlie asks his associate, P.J., to go through the files and count all the active cases containing IWRP's not signed by the client. Being even more practical than Charlie, P.J. says, "No chance, Charlie! Why don't we take a random sample instead?"

Before Charlie and P.J. can take their random sample they must decide how large a sample is necessary to insure a "reasonably precise" estimate of the proportion, P. What is "reasonably precise?" Suppose each case is classified as either IWRP signed or IWRP unsigned. Charlie must specify some margin of error, ME, in the estimated proportion, p, of cases with unsigned IWRP's that he is willing to accept. Additionally, he must specify some small risk, α, which he is willing to incur that the actual error is in fact larger than ME. That is, Charlie must specify that he wants the error in his estimate to be larger than ME with small probability α. Since an unsigned IWRP constitutes noncompliance with agency policy, Charlie wants to be fairly sure that his estimate is accurate, thus he states that he wants the probability to be less than $\alpha = .05$ that his estimate, p, differs from the **true** proportion, P, by more than ME = .05. That is, he wants his estimate to be within .05 of the true proportion 95 times out of 100.

Unfortunately, even after going through all of this specification of risk and margin of error Charlie still cannot estimate the sample size required to provide an estimate within his specified margin of error with probability $(1-\alpha) = .95$. To do this, he must be able to supply some initial **guess** as to the proportion of IWRP's which are unsigned. Suppose his estimate is denoted by $\hat{P}$. Then the desired sample size is computed from the following expression:

$$n = \frac{(z_{\alpha/2})^2 \hat{P}(1-\hat{P})}{ME^2},$$

where $z_{\alpha/2}$ is the number on the horizontal axis of the standard normal curve corresponding to an area of $\alpha/2$ to the right.

P.J. found it helpful to make the following summary of concepts which are represented by various notational devices.

P	Upper case P is used to denote the true proportion of cases in the district office with unsigned IWRP's. Without complete enumeration this number will never be exactly known.
p	Lower case p is used to represent the **estimate** which Charlie will eventually make of the proportion of unsigned IWRP's. This value of p will be computed by simply counting the number of unsigned IWRP's **in the sample** and dividing by the sample size.
n	Lower case n is the size of the sample. Clearly this number is dependent upon the degree of precision required for the estimate, p. The more accurate p must be the larger the sample must be. The current problem is to estimate the necessary sample size required to achieve a certain specified level of precision.
ME	ME or margin of error composes half of the specification of precision. The ME tells the amount by which Charlie is willing to miss the true proportion. Clearly, the larger the margin of error chosen, the smaller will be the required sample size.
α	The greek letter alpha (α) is the second half of the specification of precision. Unless the

margin of error is 1 (Charlie will tolerate **any** estimate), it is impossible to estimate the true proportion within a specified margin of error 100% of the time. Charlie must be willing to accept an estimate that is within his margin of error say 95% of the time. α is the probability that the estimate differs from the true value by more than the margin of error. Clearly, the smaller α is required to be, the larger the sample must be.

$Z_{\alpha/2}$ $Z_{\alpha/2}$ is the symbol for the $(1-\alpha/2)$th percentile of the standard normal distribution. For a given $\alpha, Z\alpha/2$ is found in a table of critical values for the standard normal distribution.

$\hat{P}$ Upper case $\hat{P}$ (P "hat") is a guess which Charlie makes as to the actual value of P. This guess is made solely for the purpose of estimating the required sample size. When in doubt, Charlie guesses $\hat{P} = .50$ since this value results in the largest sample size for a specified precision (ME and α).

Charlie believes that close to 10% of the IWRP's will not have client signatures so he gives this information to P.J. who proceeds to estimate the required sample size.

P.J.'s Steps to Estimating the Sample Size

1. Specify the margin of error ME which Charlie is willing to accept. Charlie sets ME at .05.
2. Specify the small probability α with which he is willing to exceed this margin of error. Charlie sets α at .05.
3. Find $z_{\alpha/2}$ in a table of the standard normal distribution. Selected values are given below:

$\alpha =$	.01	.05	.10
$Z_{\alpha/2} =$	2.57	1.96	1.65

4. Guess the true value of P and denote the estimate by $\hat{P}$. Charlie has guessed $\hat{P}$ at 10% or .1. A guess of $\hat{P} = .5$ will usually overestimate the necessary sample size and hence is a safe estimate.
5. Compute the required sample size using ME, $z\alpha/2$ and $\hat{P}$ from the steps above. For $\alpha = .05$, ME $= .05$, and $\hat{P} = .1$,

$$n = \frac{(1.96)^2 (.1) (.9)}{(.05)^2} = 138.$$

Table 1 on the next page gives sample sizes for various combinations of $\hat{P}$, α, and ME. It is important to note that sample sizes found by using this formula are accurate only when the population from which the sample is to be taken is very large relative to the sample size. That is, the quantity n/N should be negligible (i.e., less than .05), where N is the size of the population sampled. If, in fact, n/N is not negligible, then a **smaller** sample of size n′ may be used when n′ is obtained from n using the following formula.

$$n' = \frac{n}{1 + (n/N)}.$$

Table 1

Sample sizes for estimating proportions from large populations

			$\hat{P}$	
		.1	.25	.50
ME=.01	α=.01	5940	12375	16500
	α=.05	3456	7200	9600
	α=.10	2448	5100	6800
ME=.05	α=.01	238	496	660
	α=.05	138	288	384
	α=.10	98	204	272
ME=.10	α=.01	59	124	165
	α=.05	35	72	96
	α=.10	24	51	68

Review

Suppose you are asked to estimate the proportion of referrals to a given district office that were not contacted within two weeks of the referral date. The acceptable margin of error is ME = .01 which may be exceeded with probability $\alpha = .05$. If you guess that the true proportion is around .1 and assume the population is large, how large a sample must you take?

Answer: 3456

Alternatively, suppose that you know that there were only 10,000 referrals to that office during the period of interest. What would be the appropriate sample size?

Answer: $n = \frac{3456}{1 + 3456/10000} = 2568$

Worksheet

Sample Size When Estimating A Proportion

1. Specify the largest margin of error (ME) you are willing to accept in estimating the proportion: ME = ______________________________.

2. Specify the small probability, α, you are willing to accept that the estimate will differ from the true proportion by more than the specified margin of error ME: α = ______________.

3. Enter the table below in the column corresponding to the α selected in step 2 and find the appropriate standard normal z-score, $z_{\alpha/2}$: $z_{\alpha/2}$ = ______________________.

α	.1	.05	.01	.001
$z_{\alpha/2}$	1.64	1.96	2.58	3.29

4. Make an initial guess, $\hat{P}$, of the true proportion to be estimated: $\hat{P}$ = ______________________________.

5. Compute the preliminary estimate, n, of the required sample size using the following formula and the information from steps 1 through 4.

$$n = \frac{(z_{\alpha/2})^2 \hat{P}(1 - \hat{P})}{ME^2}$$

$$n = \frac{(\quad\quad)^2 (\quad\quad)(1 - \quad\quad)}{(\quad\quad)^2}$$

$$n = \underline{\qquad\qquad\qquad}.$$

6. If the size (N) of the population from which the sample is drawn is finite, then a smaller sample of size n' will suffice. n' is given by the following formula:

$$n' = \frac{n}{(1 + n/N)}$$

(a) Determine the size of the total population, N = ______________________.

(b) Compute $\frac{n}{N}$ using the n from step 5: $\frac{n}{N}$ = ______________________.

(c) Compute (1 + n / N): 1 + n / N = __________.

(d) Divide n by (1 + n / N) to get

$$n' = \frac{n}{1 + \frac{n}{N}} = \underline{\qquad\qquad}.$$

Unit II

Module Eight

Sample Size for Estimating a Mean

Charlie's supervisor has expressed concern that some referrals to a specific district office are not being contacted as promptly as is desirable. Thus, Charlie is given the assignment of estimating the mean time from referral to initial contact by the counselor. As with his estimation of a proportion, Charlie specifies a margin of error ME, for his estimate which he is willing to exceed with small probability α. That is, Charlie is willing to miss the true mean by as much as ME with probability $1-\alpha$ and miss the true value by more than ME with probability α.

In addition to ME and α, Charlie must also have some idea of the variability of the population to be sampled. That is, he must be able to make some guess as to the population variance. Since no prior study has been done which might provide the basis for such a guess, Charlie and P.J. conduct a small pilot study consisting of 15-20 cases to obtain an advance estimate of population variance.

From the results of the pilot study, Charlie is ready to estimate the necessary sample size using the formula below:

$$n = \frac{(z_{\alpha/2})^2 sd^2}{ME^2}$$

where $z_{\alpha/2}$ is the point on the standard normal curve leaving an area of $\alpha/2$ to the right and sd^2 is the estimate of population variance obtained from the pilot study. As before, in sampling for a proportion, this estimate may be reduced slightly if the quantity n/N is not negligible (N is the size of the population sampled). In this case the smaller sample n' may be obtained by using the following formula:

$$n' = \frac{n}{1 + (n/N)}$$

Charlie's Pilot Sample

Charlie and P.J. review a random sample of 15 case files from a district office to determine the number of days from referral to initial counselor contact. From these data, they computed an initial estimate SD^2 of the population variance.

Pilot Sample Computation

x_i = No. of Days	$(x_i - \bar{x})^2$
20	260.18
0	14.98
1	8.23
2	3.49
1	8.24
5	1.28
2	3.50
3	.76
2	3.50
1	8.24
4	.02
7	9.80
0	37.58
0	14.98
10	14.98

$$\sum_{i=1}^{15} x_i = 58 \qquad \sum_{i=1}^{15} (x_i - \bar{x})^2 = 389.76$$

$$\bar{x} = \frac{58}{15} = 3.87$$

$$sd^2 = \frac{\sum_{i=1}^{15} (x_i - \bar{x})^2}{14} = \frac{389.76}{14} = 27.84$$

Suppose that Charlie wants his estimate to be within 2 days of the true mean 95% of the time. This means that ME = 2 and $\alpha = .05$ (since his margin of error may be exceeded with probability 1-.95 = .05). The sample size required to estimate the mean with this precision may then be computed using the formula given above:

$$n = \frac{(z_{\alpha/2})^2 sd^2}{ME^2}$$

$$= \frac{(1.96)^2 (27.84)}{(2)^2}$$

$$= 26.74$$

$$\approx 27$$

This means that Charlie must review 4 more cases in addition to his pilot sample. If the total number of cases in the population sampled is small (say 200) the total required sample size may be reduced slightly by using the second formula:

$$n' = \frac{n}{1+n/N}$$

$$= \frac{27}{1 + 27/200}$$

$$= 23.79$$

$$\approx 24$$

Review

1. In addition to specifying a margin of error and the small probability with which he is willing to exceed that error, Charlie must also have an estimate of the population ______________

   ____

   before he can compute the required sample size to insure the specified precision.

2. Two ways to get an advance estimate of the population variance are (choose 1):

   a. Trial and error
   b. Pilot study or previous investigations
   c. Fortune teller or S.W.A.G.
   d. Almanac or Code of Federal Regulations

3. As the population variance increases, the sample size required for an estimate of specified precision

   a. increases.
   b. decreases.
   c. remains the same.
   d. can go up or down.

Worksheet

Sample Size for Estimating a Mean

1. Specify the largest margin of error (ME) you are willing to accept in estimating a mean: ME = ____.

2. Specify the small probability, α, you are willing to accept that the estimate will differ from the true mean by more than the specified margin of error, ME: α = ________________.

3. Enter the table below in the column corresponding to the α selected in step 2 and find the appropriate standard normal z-score, $z_{\alpha/2}$: $z_{\alpha/2}$ = ________________________.

	α			
	.1	.05	.01	.001
$z_{\alpha/2}$	1.64	1.96	2.58	3.29

4. Enter an initial guess of the population variance, SD^2: SD^2 = ________________.

5. Compute the preliminary estimate, n, of the required sample size using the following formula and the information from steps 1 through 4.

$$n = \frac{(z_{\alpha/2})^2 \, SD^2}{ME^2}$$

$$n = \frac{(\qquad)^2(\qquad)}{(\qquad)^2}$$

n = ____________.

6. If the size (N) of the population from which the sample is drawn is **finite**, then a **smaller** sample of size n′ will suffice. n′ is given by the following formula:

$$n' = \frac{n}{1+\frac{n}{N}}$$

(a) Determine the size of the total population, N = ____________.

(b) Compute $\frac{n}{N}$ using n from step 5: $\frac{n}{N}$ = ________________.

(c) Compute $(1 + \frac{n}{N})$: $1 + \frac{n}{N}$ = ____________.

(d) Divide n by $(1 + \frac{n}{N})$ to get $n' = \frac{n}{1+\frac{n}{N}}$ = ____________.

Unit II

Module Nine

The Pilot Study

Program evaluators are often called upon to conduct evaluation research requiring some type of survey. The main point of this module is to point out some of the reasons that a pilot study may be of advantage prior to initiation of a large scale survey.

It was decreed from above that Charlie should provide his superiors with an estimate of the total number of disabled persons eligible for rehabilitation services who live in a certain town being considered for the location of a new experimental rehabilitation office. Charlie, therefore, dutifully began planning a small survey. As part of his study design, he included a small pilot survey. Upon submitting his design for approval, Charlie was questioned as to the necessity of spending money on a pilot study. Having anticipated such a response from those tight-fisted guys that control expenditure of state funds, he briefly outlined the advantages of a pilot study.

I. A pilot study may be used to obtain an advance estimate of the population variance.

As was shown in a previous module, advance estimates of the population variance are necessary for the estimation of required sample size and hence for an estimate of the cost of the survey. It may be that a survey of useable precision is economically infeasible so that a major revision of goals and objectives is required.

II. A pilot study can provide valuable information regarding the survey instrument.

A pilot survey can reveal unexpected fallacies in the survey instrument such as sensitive items, nonresponse rate, inappropriate length, or item ambiguity. If a survey instrument is being used for the first time, a pilot test is highly advisable simply for working out the bugs in the instrument. The key to a successful sample survey is **planning**.

III. A pilot study can provide valuable information regarding the sampling plan.

One of the principal reasons for stratifying in a sample survey is that the sampling problems encountered may differ greatly from one segment of the population to another. For example, sampling of apartment dwellers presents different problems from sampling of single family suburban dwellers. A properly designed pilot survey can often point out specific sampling problems which may be encountered and suggest appropriate methods of stratification. Additionally, if a pilot study indicates heterogeneity among various subpopulations, stratification may be used to achieve greater precision at reduced cost.

Upon completion of Charlie's presentation, the conference room was filled with deafening applause, and Charlie was dubbed hero of the day for bringing this new enlightenment to the program evaluation section.

Review

Check the valid reasons for conducting a pilot study prior to a sample survey.

1. To provide information for the estimation of survey cost. ______
2. To become a hero at the office. ______
3. To improve the survey instrument. ______
4. To run up the cost of the survey. ______
5. To improve the sampling plan. ______
6. To increase the credibility of results. ______

Unit II

Module Ten

Sample Size with Several Sub-populations or Several Items

While reviewing case files on Friday afternoon, P.J. began to realize that conducting a separate case review type survey for every single item of interest was grossly inefficient. Thus, he suggested to Charlie that they review one sample of cases for a group of items such as:

1. The proportion of IWRP's which have not been signed by the client;
2. The mean time for referral to initial contact by the counselor;
3. The proportion of IWRP's which have been amended;
4. The mean time required to receive the evaluation report after the end of evaluation.

Further, he asks if it is necessary to design a separate sample for each district office or service unit of interest. "A good suggestion and a good question," replied Charlie, "Now let's find out how to determine our sample size."

After several hours of research and contemplation, Charlie makes the following observations:

1. If there are a moderate number of items and the desired precision (ME and α) is specified, it may be possible to calculate the required sample size for each item and use the largest of the sample sizes found.
2. If it is infeasible to calculate a sample size for each item, or if the largest of all the sample sizes calculated is economically infeasible, it may be possible to specify the **most important** item in the survey and use the sample size necessary to insure the desired precision for that item.
3. If estimates are to be made separately for several subdivisions of a given population (e.g., the district office or administrative areas in a state rehabilitation agency) then a separate calculation of sample size is necessary for each subdivision. If the item variance and precision requirement are the same for each of 10 subdivisions, for example, then the total required sample size will be nearly 10 times as large as that required for a single estimate for the whole population. That is, the ability to make precise estimates about population subdivisions is relatively expensive in terms of increased sample size.

P.J. comments, "I had no idea that just adding a few subdivisions to our study while retaining the same desired level of precision could be so expensive. Maybe we should rethink our precision requirements and the actual utility of the subdivision estimates."

"Good thinking," says Charlie.

Review

1. When several items are included in a study and required sample sizes may be computed for each item, the actual study sample size should be the ____________ item sample size.
   a. largest b. smallest

2. If precision requirements (α and ME) are the same for two items, the one with the largest variance will require the ____________ sample size. a. largest b. smallest

3. It is sometimes appropriate to determine the ______________ item in a survey and use a sample large enough to insure adequate precision of the estimate for that item.
   a. most important
   b. least important
   c. most controversial
   d. least variable

4. Separate estimates for population subdivisions are ______________ in terms of overall sample size.
   a. expensive
   b. cheap
   c. infeasible
   d. inaccurate

66

References

Selby, S.M., (Ed.). **Standard Mathematical Tables**. Cleveland, Ohio: The Chemical Rubber Company, 1969.

Unit III

Concepts in Measurement

68

Table of Contents

Module **Page**

Levels of Measurement 75
Objective: To introduce and define the different levels of measurement

Measurement Standardization and Frequency Distributions 79
Objective: To introduce the concepts of percentile ranks, histogram and frequency distribution

Computing Standard Scores and Derived Scores 87
Objective: To illustrate the use and computation of percentile ranks and standardized Z-scores

Reliability 91
Objective: To introduce the concepts of reliability and discuss the various approaches to estimating reliability

Validity 95
Objective: To introduce the concept of validity and discuss the various types of validity and methods of validity assessment

Sources of Data in Vocational Rehabilitation 101
Objective: To discuss the various types of data (such as rating scales) which are available to program evaluators

Unit Three

Module One

Levels of Measurement

Throughout time people have sought better ways to group different entities so that they may be better understood. Measurement refers to different techniques by which objects can be grouped and interpreted. Generally the terms **measurement** and **quantification** are interchangeable. If something exists, or is hypothesized to exist, it can be assigned a number that serves as a reference point for further analysis. Measurement techniques enable us to agree on **how much** of an entity exists. In the area of human services, the measurement of psychological well-being, life satisfaction or self-concept is much less precise than the measurement of money or time spent in a program. However, any measurement scheme, assuming certain rules are followed, results in increased objectivity.

> **Measurement is a process by which evaluation information is converted into data which can be objectively interpreted.**

There are four different levels of measurement, each of which permits increasingly potent types of statistical analysis. These varying levels of measurement (measurement scales) require different assumptions. In general, the more assumptions a measurement scale meets, the more potent, in terms of drawing inference, it is.

The simplest type of quantification, the **Nominal Scale**, classifies people or objects. Numbers or symbols are used to identify the groups to which various persons or objects belong. Categorizing persons by disability type is an example of the nominal measurement scale in action.

The second level of measurement is the **Ordinal Scale**. Ordinal scales have an arbitrary zero point and the differences between scale intervals are not truly equal. For example, client self ratings of their satisfaction with services are often measured by an ordinal scale, as illustrated below.

How satisfied were you with rehabilitation services?

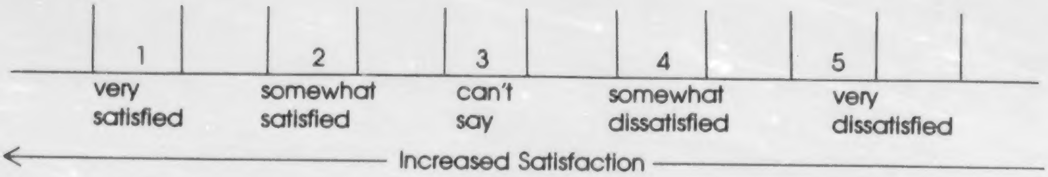


Figure 1
Example of an Ordinal Scale

The actual increase in satisfaction from "somewhat satisfied" to "very satisfied" might be considerably greater than the increase from "somewhat dissatisfied" to "can't say" although each change represents a gain of 1 "unit." Hence, strictly speaking, the difference between intervals 3 and 4 would not be the same as between intervals 1 and 2. Note that there is no true zero point, i.e., no point where there is zero degree of satisfaction. The advantages of the ordinal scale are that it may be used to rank persons with respect to specified stimuli, rather than on the basis of purely subjective judgments.

The **Interval Scale** requires that the scale intervals be equal, but, like ordinal measurement, does not require a true zero point. Some psychological measures, such as IQ tests, approach an interval scale. For example, on certain IQ tests the difference between IQs of 50 and 60 is the same as the difference between IQs of 110 and 120. However, because the zero point on an IQ scale is merely abritrary (there is no such thing as zero intelligence), IQ scales are interval not ratio in nature.

The common bathroom scale has the properties of a **Ratio Scale**, which is the fourth and highest level of measurement. In order to qualify as a ratio scale the entity being quantified must be capable of not existing, that is, the scale must have an absolute zero point (e.g., no weight at all). Second, the scale intervals must be equal. On the bathroom scale the difference between 10 and 11 pounds is the same as the difference between 110 and 111 pounds. In evaluation research, time, money and various physical variables such as height and weight are quantified via a ratio scale. Seldom do psychological tests, rating scales or outcome measurement scales represent ratio scales.

Most of the social psychological data encountered in program evaluation research is in the form of ordinal data. There are limitations on ordinal scale data in either describing the population under study or in drawing inferences for program improvement. Actually many ordinal scales, as they approach interval scales, are analyzable via more powerful statistical tests, as long as other underlying statistical assumptions are not violated. The rigorous statistician may prefer that the raw ordinal scale data be first converted, by appropriate transformations, to interval scale data or that less powerful, but entirely appropriate nonparametric statistical techniques be used (see nonparametric statistics, Unit 4) in the analysis of nominal and ordinal data.

Unit Three

Module One

Review

1. Measurement can be defined as:

2. Name the different levels of measurement and give a description of each

a. Name:

Description:

b. Name:

Description:

c. Name:

Description:

d. Name:

Description:

3. Discussion Question: Why is it important to attend to the type of measurement used in an evaluation scheme?

Unit Three

Module Two

Measurement Standardization and Frequency Distributions

When a program evaluator uses data that fits a ratio measurement scale, such as money in a cost analysis study, there is no problem with measurement interpretation. As the evaluator moves to more person-oriented behavioral attributes, precision in quantifying client characteristics becomes more difficult to achieve and measurement must be done at the ordinal or interval level. In addition to measurement problems, these behavioral measurements often present problems of interpretation. For example, we know what it means to have spent $1000 on a client's vocational training, but what does it mean for a client to score 1.5 on a self-concept scale? To enhance the interpretation of behavioral measurements, it is often possible to construct different measurement devices (tests, inventories, rating scales, etc.) and convert the raw scores to group centered scores. This process is called **standardization** or norming of measurement scores. Standardization allows any one individual's score to be interpreted by comparison to the scores of a group of individuals.

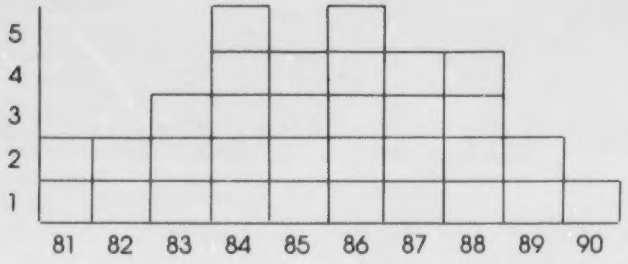
One type of standardization process converts raw score data to percentile ranks. That is, the scores are converted to a scale of 1 to 99. A person's score converted to a percentile of 75 means that based on the norm group the person scored higher than 75 percent of the group and lower than about 25 percent of the group, i.e., the person is in the top 25%. The 50th percentile divides the group into equal halves. Publishers of psychological tests often have converted raw scores to standardized scores. When using these tests it is a simple matter to convert raw scores to percentile scores by referring to the conversion tables provided by these publishers. Often score conversions are given for different age groups, geographical regions or groups of individuals. For example, the **Human Service Scale**, a measure of rehabilitation client needs, has conversion tables which correspond to the different vocational rehabilitation disability groups.

Conversion of raw test scores to percentile ranks permits definition of the individual's score in terms of specific reference groups.

In evaluating a program it is not unlikely for the evaluator to be asked to formulate some specialized kind of measurement device in order to assess some program specific variable. Or, an administrator may wish to establish normative data for different program units. In both cases the evaluator would need to know how to formulate derived scale scores. In order to understand the computation and meaning of derived scores it is necessary to have some understanding of frequency distributions, the normal curve, and such descriptive statistics as the mode, median and standard deviation.

Assume that we have constructed a test, say the **Test of Community Rehabilitation Attitudes [TCRA]**, and have a set of raw scores. The scores are merely numbers and in themselves lack meaning. One way to give some meaning to the scores is to plot a histogram of raw scores for those persons who competed the test.

73



TCRA Scores
Figure 1

Histogram of Raw TCRA Scores for 32 Persons

Figure one is a **histogram** or frequency distribution of the test scores for persons who completed the TCRA. Each square represents one person and each number along the baseline represents a scale score interval one half a point above and below the actual score. By counting the number of squares per score the relative distribution of scores is readily apparent and cumulative frequencies are ascertained.

Cumulative frequencies represent the number of people receiving the same or lower test score. For example, in Figure 1, 12 persons received a score of 84.5 or less, 21 persons a score of 86.5 or less.

In conducting evaluation research it is often desirable to describe a population of clients relative to a variable such as weekly earnings at referral. The frequency histogram is a technique which can make such distributions more understandable. Suppose an evaluator has a sample of 100 clients and wishes to describe their weekly earnings at referral. To do this using a frequency histogram one must perform the following steps:

1. Find the smallest and largest weekly earnings and compute the **range** by subtracting the smallest from the largest. (If $10 is the smallest and $160 is the largest weekly earnings, the range is $150).

2. Divide the range into five to fifteen equal intervals. Clearly the **number** of intervals will determine the **size** of the intervals. The size of the interval may be computed by dividing the range by the number of intervals chosen. If we choose 10 intervals, they must be of size $15.

3. Write down the intervals in a vertical line as follows:

$10 - $24
$25 - $39
$40 - $54
$55 - $69
$70 - $84
$85 - $99
$100 - $114
$115 - $129
$130 - $144
$145 - $159

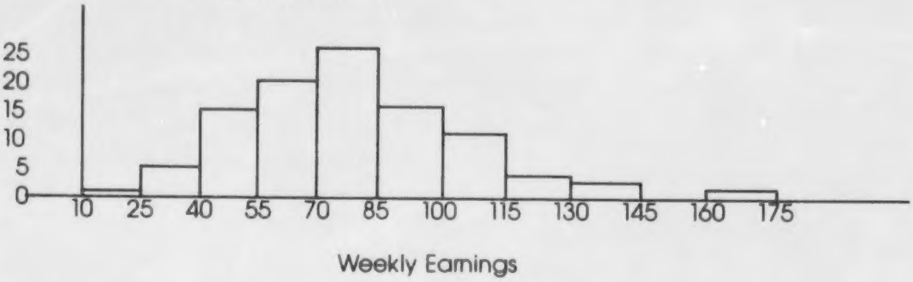
Since the last interval does not include the largest observation $160 we must add an additional interval, $160 - $174.

74

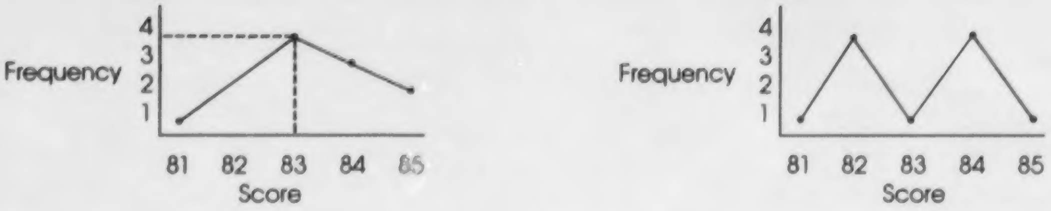
4. Count the number of observations (weekly earnings that fall within the defined intervals) and list the frequency beside each interval as follows:

Interval	Frequency
$10 - $24	2
$25 - $39	5
$40 - $54	15
$55 - $69	20
$70 - $84	25
$85 - $99	15
$100 - $114	10
$115 - $129	4
$130 - $144	3
$145 - $159	0
$160 - $174	1
	100

5. Plot the weekly earnings intervals along the horizontal axis of a bar graph and the frequencies on the vertical axis as follows:



Frequency polygons can be generated by hand or through a computer. These frequencies are plotted on a graph similar to the histogram, but using dots to represent the number of persons per scale interval mid point. When used in conjunction with statistical measures of central tendency, these distributions facilitate score interpretations (both within groups and in comparison to other groups of people taking the same test). For example, Figure 2 represents the score distributions for two different groups of persons who completed the TCRA. Cumulative frequencies (the number of persons scoring less than or equal to a specified number) and the number of individuals receiving any one score can be easily ascertained from the frequency polygons in Figure 2. In group A, there were, for example, four individuals who had a score of 83, and five individuals who had scores of 83 or less. As illustrated by Figure 2, the two groups have different score distributions. Specifically, the number of persons in each group receiving any particular score may be quite different. Measures of central tendency can help to describe different distribution of scores.



Group A

Raw TCRA Scores for 10 persons

Group B

Raw TCRA Scores for 11 persons

Figure 2

Cumulative Score Frequencies for Two Groups of Persons Completing the TCRA

Four statistics, the mean, median, mode and standard deviation, help in interpreting sample scores. The sample **mean**, often noted as $\overline{X}$, (X bar), is simply the arithmetic average for the population measured. It is computed by summing the scores and dividing by the number of persons taking the test. The mean for group A, Figure 2, is 82.7, the sum of the scores (827) divided by the number of persons (10). For group B, the mean score equals 83.0. The sample **median** is that score at which half the persons score higher and half lower and is identical to the 50th percentile. From Figure 2 it can be seen that although the distributions are markedly different, both have medians of 83. The median is a good aid to interpretation when a distribution is skewed in one direction. Such is often the case when income levels are considered. Usually in a sample of rehabilitation clients, incomes will pile up at the low end of the distribution with a few individuals falling in the high end. Assume for example, that 20 persons earn $60 a week and five persons earn $200 a week. Using the mean as an indicator of central tendency would distort the data to make it look like the entire group had higher earnings (mean of $80 vs. median of $60) than it actually did. In this case the median would be more representative of group earnings. The sample **mode** is defined as the most commonly occurring score in a distribution. In group A the mode is 83. Group B is bimodal; the most commonly occurring scores are 82 and 84. Use of the mode can be illustrated by comparing groups A and B. Both groups have similar mean scores but the mode indicates that the distributions are different and the groups may not be comparable.

> **The mean is the average score for a group, the median is that score occurring at the 50th percentile, and the mode is the score that occurs most often. In the normal curve the mean, median and mode occur at the same spot on the baseline.**

Before considering computation of derived scores one final descriptive statistic should be described. Measures of **variability** describe the relative range of scores about the mean. The most widely used measure of variability is the sample standard deviation. The standard deviation, SD, is used as the standard unit appearing on the baseline underneath the density curve of a distribution of scores. Figure 3 illustrates the normal density curve marked off in standard deviation units (SD).

The shape of frequency distribution curves can be described by standard deviations. For example, under the normal curve, a theoretical distribution which can be completely described mathematically, 68% of a group's scores theoretically will be found within one standard deviation above and below the mean, 95% of the scores will theoretically be found within two standard deviations above and below the mean and so forth. (Note that in the normal curve, the mean, mode and the median occur at the same place – the center of the distribution).

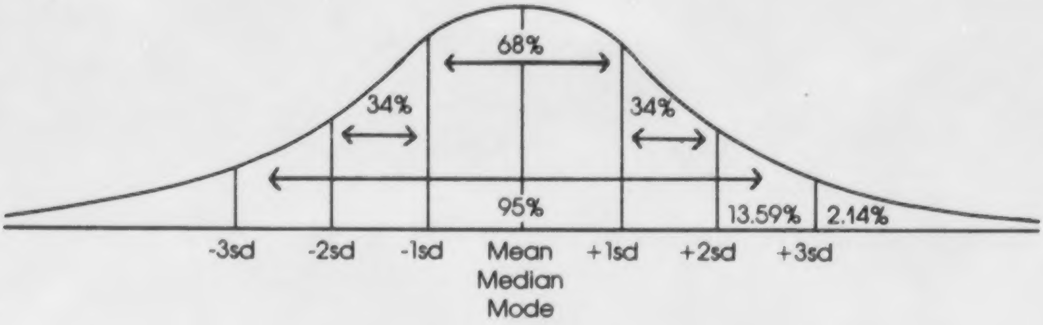


Figure 3

Illustration of the Normal Curve

Standard deviations allow for comparisons of intra and inter group scores. Standard deviations are computed as follows:

1. Compute the mean for the group of scores.
2. Subtract the mean from each of the individual scores.
3. Square each of the values obtained from step 2.
4. Sum the squared deviations obtained from step 3.
5. Divide the above sum by one less than the number of persons in the group.
6. Find the square root of the number obtained in step 5 by referring to a table available in any basic statistics book. The resulting value is the standard deviation.

The standard deviation can be expressed as $SD = \frac{\sqrt{\Sigma (x-\bar{x})^2}}{n-1}$

where:

$\sqrt{\ }$ means take the square root

Σ means sum these numbers

x is an individual raw score

$\bar{x}$ is the group mean

$(\quad)^2$ means perform the operation inside the parentheses then square the result

n is the number of persons with scores

For persons in group A, Figure 2, the standard deviation equals 1.37. Thus 68% of that group falls between TCRA scores of 84 and 81.

Unit Three

Module Two

Review

1. Standardization consists of ____

____.

2. Published tests usually convert raw score values to __________ which refer to score norms based on different ______________ of people.

3. The way to plot a distribution of scores is to graph the total number of __________ on the baseline and mark the ______________ of persons per scale score.

4. Match the following:

A. mode

B. median

C. standard deviation

D. mean

1. average of group scores
2. point at which most persons score
3. point at which 50% of the scores fall below and 50% above
4. measure of group variability

5. For the normal distribution, the area within one standard deviation of the mean will account for ___% of the units under the curve.

6. The normal curve is a ______________ distribution.

7. Given below are the ages at referral of 25 rehabilitation clients. Complete a frequency histogram using 5 intervals.

Age at referral

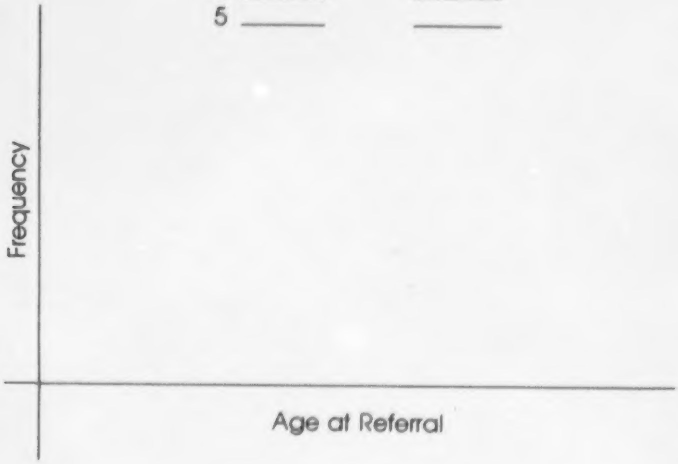
19	26	30	32	41
19	27	30	34	48
21	28	31	36	50
25	28	32	38	59
26	29	32	40	60

7. (cont.)

Range = 60 - 19 = ______

Interval Size = Range/5 = ____(round interval size to next largest whole number)

	Interval	Frequency
1	______	______
2	______	______
3	______	______
4	______	______
5	______	______



Unit Three

Module Three

Computing Standard Scores and Derived Scores

Standardization of sample scores facilitates score interpretation for individuals in a group and enables the evaluator to compare individuals across groups. Standardization of measurement units permits the evaluator to attach practical meaning to the raw numbers. By standardizing program specific evaluation instruments the evaluator can minimize interpretive errors.

Perhaps the most common way to transform raw scores is through conversion to percentile ranks. This conversion can be carried out by (1) listing the cumulative relative frequency of the scores, (2) finding the score interval midpoint, and (3) multiplying by 100 to obtain a percentile rank. Thus all scores are converted to percentiles and raw scores are interpreted on a scale from 1 to 99. A person at the 99th percentile would have scored in the top 1% of the group, a person at the 1st percentile would be in the bottom 1% of the group.

Percentile rank conversion is an effective way to interpret an individual's rank within a group.

In evaluation research one is often faced with large amounts of measurement data. In this case it is often more efficient to compute cumulative percentages (cumulative relative frequencies) rather than percentile ranks, to determine an individual's rank within a group. Cumulative percentages represent the upper boundary of the score interval. They are easily calculated by computer processing and provide the same function as percentile ranks, i.e., classifying individual scores with regard to relative position of all persons in the group. Table 1 is an example of a computer printout of cumulative frequencies on a set of test scores for a group of 57 persons.

Table One

Cumulative Percent Illustration

SAS-NUMFREQ **11:06 Thursday, April 21, 1977**

Frequencies Psychological Tests

MMPI TEST SCORE	FREQUENCY	CUMULATIVE FREQUENCY	PERCENT	CUMULATIVE PERCENT
5	1	1	1.754	1.754
11	11	12	19.298	21.053
13	6	18	10.526	31.579
15	12	30	21.053	52.632
17	8	38	14.035	66.667
18	9	47	15.789	82.456
20	5	52	8.772	91.228
22	3	55	5.263	96.491
24	1	56	1.754	98.246
25	1	57	1.754	100.000

There are several limitations in using percentile ranks or cumulative percentages. One limitation is that in normalizing raw scores small differences at the center of the distribution appear to be larger than they really are, and large differences at the distribution extremes appear to be smaller than they really are. Table 1 illustrates this phenomenon. The middle of the distribution, from cumulative percentage 31.579 to 66.667, a difference of 35 percentage points, represents persons who are only five score units apart. At the upper extreme, from cumulative percentages 82.456 to 100.00, there is a difference of only 18 percentage points yet there is an actual score difference of seven.

Another limitation in the use of percentile ranks and cumulative percentages is that averaged percentages are difficult to interpret. This limits possibilities for comparison between different groups. Comparisons between different groups can only be made by examining the score distributions.

A more effective technique for intergroup comparisons is the conversion of raw scores to **derived standard scores**. Standard scores are transformations which make use of the group mean and standard deviation. By computing standard scores, it is possible to compare scores from different measurement instruments. This is because standard scores convert the raw scores to a scale which is relative to the group's standard deviation and centered around the group's mean, regardless of the magnitude and range of raw scores.

A common standard score is the Z-score. The procedure for converting raw scores to Z-scores is as follows:

1. Take any raw score and subtract from it the group raw score mean.
2. Divide the above value by the standard deviation. This value is the Z-score.

The formula for computing a Z-score is: $Z = \frac{X-\bar{X}}{SD}$
where X = any raw score

$\bar{X}$ = the group's raw score mean
SD = the group's standard deviation

The Z-score for the five persons in Table 1 who received a raw score of 20 is Z =

$$\frac{20-15.74}{3.86} = 1.10.$$

Thus persons who received a raw score of 20 are located approximately 1.1 standard deviations above the mean (15.74) for this group. Assuming that the distribution of raw scores approximates a normal distribution, the Z-scores approximate a standard normal distribution and hence are convertible to percentile scores using tables available in any statistics text. This represents a distinct advantage in communicating measurement findings. Z-scores take either negative or positive values, corresponding to the different sides of the mean. Another kind of standard score, the T-score, is based on the Z-score and eliminates decimals and negative numbers. T-scores are computed by first computing a Z-score as above, then multiplying the Z-score by 10 and adding 50. The T-score for the raw test score 20, Table 1, is 61. If the raw scores approximate a normal distribution then the T-scores are approximately normally distributed with a mean of 50 and a standard deviation of 10.

Standard scores eliminate the distortions of percentile ranks and through use of the standard deviation define what is meant by a scale interval and a scale zero point.

Unit Three

Module Three

Review

1. The most common raw score to derived score conversion for comparing individuals within a group is ____________________.

2. Besides using percentile ranks for intra group comparisons one may use ____________________.

3. Two limitations in the use of percentile ranks or cumulative percentages are:

a)

b)

4. Standard scores are transformations which make use of the group ______________ and ______________ ______________.

5. Given raw score values of 16, 23, and 30 for individuals in a group with a mean of 24 and a standard deviation of 4.0, what are the respective Z-scores? a) __________, b) __________, c) __________

6. The main advantage of standard scores in evaluation measurement is that they ______________ and ______________.

82

Unit Three

Module Four

Reliability

The estimation of measurement reliability is a topic covered in most textbooks on research or statistical methods. Our purpose in this module is to introduce several concepts regarding reliability. Measurement theory rests on the premise that if something exists it can be accounted for in a quantifiable form. Reliability refers to consistency in measurement. Reliability is intimately related to another concept, validity (Unit Three, Module Five). Basically a test is thought to be valid when it measures what it purports to measure. However, a test cannot be valid unless it is reliable or consistently measures what it purports to measure.

Reliability or measurement consistency is thought of in terms of partitioning or splitting a person's obtained score into two parts, the true score and measurement error. Measurement error can consist of many things such as scoring errors, changes in the test taking environment, inappropriate test items, etc. Both the true score and the measurement error are present in any obtained score, and the relative size of one compared to the other defines the amount of reliability. The objective is, of course, to minimize measurement error so that the obtained scores are very nearly equal to the true scores. There is, however, no way that true scores can be isolated from simultaneously occurring measurement error. Rather, several techniques have been developed by which the level of measurement error can be estimated.

Reliability can be thought of as test score consistency and can be assessed by estimation of measurement error.

Methods of estimating reliability

Perhaps the most common way to estimate reliability is through the test-retest method. This method consists of testing a group of individuals twice on the same instrument. The basic assumption is that changes in individual scores represent measurement error. A reliability estimate is expressed as a ratio, ranging from zero (when observed test scores are completely due to error) to one, (no measurement error is present), and test re-test reliability is estimated by some measure of association, generally a product-moment correlation. Thus the degree to which a group's test scores fluctuate from test to retest is expressed as a correlation between scores over time. A test that has a reliability coefficient of .70, is more reliable than a test with a reliability coefficient of .30. The closer the decimal is to one, the more reliable the test. Because the major factor in the test-retest method is time, this type of reliability estimate may also be called a **coefficient of stability**.

Another way of estimating reliability is through the use of parallel tests. With this procedure two separate tests designed to be parallel to one another (corresponding items on each test measure identical things) are both administered to a group. With this method measurement error is the difference between an individual's scores on the two parallel forms, again expressed as a correlation coefficient. An obvious advantage to the parallel test method is that no waiting period is required before determining test reliability. However, it is often very difficult to construct two tests so that they are, in fact, parallel. Somewhat similar to parallel test reliability estimation is the split-half method. This procedure involves a comparison of scores on each half of a single test by summing and comparing the scores for odd and even numbered items. There are two assumptions which must be met before using the split-half method. First, the item pairs or test halves must be equal in terms of difficulty, content, etc. Second, individuals taking the test must have the opportunity to complete the entire test. Tests which have a speed component, such as sorting bolts in a limited time period, are not amenable to the split-half procedure. Both split-half and parallel test reliability estimates are sometimes referred to as **coefficients of equivalence** since they indicate the degree of equivalence of the parallel forms or the two test haves.

A fourth way to estimate measurement reliability is the **Kuder-Richardson formula**. This technique, referred to as K-R 20, results in an index of test reliability when the test items are scored dichotomously (yes-no, right-wrong, etc.) and all the items of the test measure a single concept. The logic of the K-R 20 technique is that in a unifactor test there should be a point at which an individual has answered correctly all the questions possible and beyond that point the questions are too difficult. The degree to which such "cut points" are established for different individuals is an estimate of the test's reliability. Again reliability is expressed as a decimal coefficient. K-R 20 reliability estimates are most often given for achievement and aptitude tests. The K-R 20 estimate may be considered a special case of, and sometimes referred to as, **coefficient Alpha**, a general measure of test internal consistency.

The above types of reliability estimates are most likely to be associated with paper and pencil, objective tests. The program evaluator is often faced with estimating the reliability of measurement devices which have a built in degree of subjectivity. Such is the case when the evaluator uses rating scales (Unit Four, Module Six). Rating scales are likely to be constructed with a program specific method in mind. When such a scale is used the evaluator can determine the measurement reliability through the **variance component method**. This procedure utilizes an analysis of variance model and partitions score variability into true score and error variance. Error variance can be defined as the residual variance that remains after accounting for rater (and/or item) variability. This procedure is somewhat complex. A simpler method for reliability estimation, especially when one is using an established rating scale, is to determine the degree of rater agreement.

Evaluating reliability estimates

All of the different forms of estimating reliability have both advantages and disadvantages. No one method is appropriate for all situations. For example, the test-retest method can be biased depending on the amount of time from test to retest. If the time is short, persons may remember how they respond at the first testing and the reliability estimate will be an overestimate of the actual reliability. On the other hand, if the time between testing is long, changes in individuals taking the test affect the test scores. Since these changes are not measurement errors, the reliability coefficient will be an underestimate of the true reliability. **Parallel test reliability** eliminates the chance that test takers will remember how they previously responded. It is considered a more conservative estimate of reliability since the true reliability is probably not lower than the estimate. The **split-half** technique has a major fault. In that split halves of a test are used in estimating reliability and because of the fact that the longer a test is the more reliable it becomes, the technique can underestimate true reliability. Fortunately a technique, called the **Spearman-Brown prophecy formula** can make up for this deficiency by statistically "expanding" the size of each test half. The reliability estimate method depends upon the kind of measurements taken. In the case of paper and pencil test reliability, the best method would be to combine the parallel test and test-retest method.

Different methods of estimating reliability have specific advantages and disadvantages. The method chosen depends on the type of measurement device being used.

Besides being aware of the fact that different estimation procedures produce different estimates of reliability, the evaluator needs to be cognizant of different factors affecting reliability estimation. The fact that test length affects reliability has already been mentioned (the longer the test, other things being equal, the more reliable it is). Time allocated for answering test items also affects the K-R 20 reliability estimate. Other important factors are group homogeneity and irregularities in administration of the measurement instrument. Group **homogeneity** occurs when persons in the group are all at the same level on whatever is being measured. Group homogeneity will reduce the reliability estimate. It is better for the group to be heterogeneous, i.e., exhibit a range of test scores.

One final issue concerns determination of what can be considered an acceptable level of reliability. This concern is an important one because the reliability level of a test places limits on the validity of the test. The relationship between measurement reliability and measurement validity is as follows. When test validity is taken as the correlation between a test and some error free criterion, maximum validity cannot be more than the square root of the reliability coefficient. Thus if a test has a reliability coefficient of .70, the validity coefficient cannot be higher than .84. Given the relationship between reliability and validity, the question of what constitutes practical minimal reliability is difficult to answer. It depends on the degree of precision one will accept, the type of measurement instrument used and the purpose for using the instrument. Different types of published tests exhibit different ranges of reliability, for example, attitude inventories tend to be less reliable than achievement tests. Sometimes a check on the available literature and reported reliability can establish the generally accepted level for a type of measurement instrument. For example, rating scales which are often situation specific, generally have reliabilities ranging from .60 to .90.

Finally the purpose for using a measurement instrument influences what is considered an acceptable level of reliability. If an instrument is used to make diagnostic decisions such as the determination of a client's treatment program, the test should be highly reliable, one with a reliability coefficient in the .90's. On the other hand, an evaluator who is trying out an experimental measurement instrument applied to group data could accept a somewhat lower reliability coefficient.

Unit Three

Module Four

Review

1. Reliability consists of partitioning a score into two parts, the ________ ________ and ________ ________.

2. The reliability estimate which can be considered a measure of test internal consistency is ________.

3. Match the following:

   1. test-retest
   2. parallel test reliability
   3. split-half reliability
   4. Spearman-Brown prophecy formula

   a. sets lower level of true reliability

   b. can underestimate or overestimate reliability

   c. used to adjust split half estimates

   d. can underestimate reliability

4. What are three factors that affect reliability estimation?

   a.

   b.

   c.

5. The following information is available on the WAT, a test that measures rehabilitation client workshop ability.

   Norm sample of 125 clients, parallel form reliability of .73, test-retest on parallel forms (six weeks from test to retest) reliability .89, coefficient Alpha equals .53.

   What are some of the possible implications from the above information? What additional information might be valuable?

Unit Three

Module Five

Validity

The program evaluator is always concerned with the validity or truthfulness of reported figures. In vocational rehabilitation program evaluators often report the degree of agency success as the number of persons who are employed following services. Employment is assumed to be a valid measure of rehabilitation. Client employment is generally accepted as the system wide measure of program success. Similarly, client earnings and employment are criterion measures or measures of ultimate outcome. Criterion measures are sometimes so obviously valid as to preclude further discussion, for example post service employment most certainly represents vocational rehabilitation. However, the validity or realness of such criterion measures can, and has, been questioned. For example, Kunce, Miller, and Cope (1974) studied the vocational rehabilitation system on a nationwide basis. They found that average rehabilitant salary at closure and the number of persons rehabilitated (i.e., employed at closure) were inversely related. In other words, an agency might have high rehabilitation rates, but the rehabilitants would have relatively low salaries. Another agency could have just the opposite kind of outcomes. If only one criterion is chosen as the measure of agency success, some agencies might be judged as being less effective. Assuming that employment and salary are both important outcomes, some agencies might choose the "wrong" criterion. For those agencies, that inaccurate criterion would not be a wholly valid measure of service effectiveness.

Another example of questionable validity occurs when there are constant errors in reporting. A state agency might conduct an audit of counselors reporting "26" closures. Say the audit revealed that 32% of clients reported employed at closure were, for one reason or another, in fact not employed. If the errors in reporting were constant across counselors and reporting periods the reports would be in fact reliable, but for agency purposes not valid. This last example illustrates another important point, validity is always relative. To judge whether the 68% accurate closure rate for the agency above is good or bad, we would need to know the corresponding figure for other similar agencies.

Test validity

Validity concerns become even more important for measurement of client characteristics not based on direct behavioral criteria. There are times when the evaluator is concerned with determining if client psychosocial variables are related to program process characteristics or client outcome. The client characteristics are inferred from various tests, and the validity, or the degree to which the instruments measure what they purport to measure, becomes very important. Obviously a measure can be valid for one purpose and not another.

> **Validity is defined as the degree to which a measurement instrument assesses what it is designed to measure.**

As mentioned in Module Four the concept of reliability is closely linked to the concept of validity. A measurement device must be reliable before it can be valid, although reliability does not automatically imply validity. For example, a bathroom scale may consistently register a certain weight even though that weight is not a true representation of a person's weight. Reliability estimation is concerned with variable and random errors while validity estimation deals with constant errors. As in the estimation of test reliability, validity estimates are determined by several different procedures.

A test is never completely "valid." It may be highly valid for one purpose, somewhat valid for another purpose and completely invalid for a third purpose.

Validity is sometimes expressed like reliability, in terms of a coefficient or index ranging from zero to one. The way the coefficent is obtained varies according to the validity estimation procedure utilized. The most common type of validity is **face validity**. Face validity is a subjective appraisal of the degree to which a test appears to measure what it is designed to measure. A test that measures anxiety should have items that look like they measure anxiety. Face validity is a very crude technique, and it is not indexed.

Content validity is somewhat like face validity in that it deals with test items. However, establishing content validity is a much more rigorous procedure. One way to estimate content validity is to carefully determine what items make up the domain or content of some specific behaviors. Because there can be hundreds of such items for most behaviors, the evaluator draws a sample of items which are thought to represent different levels of the behavior. When the items are pooled together it is thought that the resulting test will discriminate between individuals at different levels of mastery. A statistical procedure, called factor analysis is sometimes used to assess content validity. This procedure groups test items together into different factors, suggesting that the items in a factor relate to some specific content or knowledge base. Measures of internal consistency also are used to express the degree to which a test measures a specific form of content. Content validity is most likely to be reported for achievement tests and is not usually indexed.

Predictive validity. This refers to the relationship between a test and some criterion. This relationship is usually indexed by the correlation between test scores and the criterion. Predictive validity is established empirically, i.e., the test is given to a group of persons and the eventual outcomes for the group are ascertained. If the test has predictive validity there will be a relationship between the scores and outcome. Presumably individuals with higher test scores will have better outcomes, individuals with lower test scores, poorer outcomes. The idea of using a test with predictive validity is that the evaluator can predict, within known limits, what groups of people need service interventions in order to reach a desirable outcome. In fact, it is possible to devise an equation, using several tests, that can index an individual's expected level of success.

Predictive validity assumes that the test is first constructed, then the relationship with the criterion is established through a followup study. This is a rather time consuming process. **Concurrent validity** is related to predictive validity, except that the relationship between test and criterion is established simultaneously. A group of persons who have not experienced the criterion behavior take the test and their scores are correlated with the scores of persons who have experienced the criterion behavior. The Minnesota Multiphasic Personality Inventory is an example of a test originally established through concurrent validity. Here the score levels between persons who had received various psychiatric diagnoses were compared with scores of "normals." The difference in the scores was taken as a measure of psychological abnormality. Predictive and concurrent validity may be referred to as **empirical validity.** Predictive and concurrent validity are indexed by the correlation between test and criterion.

Construct validity is another type of validity estimation. It is more theoretically involved than the other types and is most likely to be used by the evaluator who is interested in measuring psychological traits and characteristics. Actually, construct validity underlies all forms of validity, but simply defined, it refers to the degree a test measures any hypothetical construct. For example, anxiety is a hypothetical construct which can exist in different forms. What anxiety is, where and when it is likely to be experienced is linked to psychological theory. The idea behind construct validity is to suggest that something exists, relate it to theory concerning the phenomena, predict how a test purported to measure the construct would differentiate people in terms of the construct, and finally, empirically check the predicted differentiation. Sometimes investigators correlate the results of a new test with scores of an older test that is generally accepted to measure the construct in question. The level of association is then taken as an index of construct validity.

Validity estimation depends on the purpose of the measurement instrument. Content validity is used for measures of achievement, predictive and concurrent validity relate to actual outcome behaviors, construct validity is used to relate theory to practice.

Evaluating Validity Estimates

As with reliability estimation there are several different factors to be considered in evaluating and interpreting validity estimates. By now it should be apparent that validity estimates are not absolute. The meaning attached to any estimation must involve the purpose of the test.

Evaluating empirical validity. Generally the higher the validity coefficient (i.e., the relationship between test scores and outcome) the better the test. One of the most important considerations is the type of criterion used to establish empirical validity. An evaluator interested in predicting work adjustment would be more interested in a test, everything else being equal, that correlated .70 with satisfactory placement, than in a test that correlated .85 with attendance at group therapy. The evaluator should also be interested in the similarity of the reported criterion group to the group of current interest. A test which is valid in predicting executive work behavior may not be valid in predicting the work behavior of rehabilitation clients.

In assessing the meaning of a measurement instrument's predictive validity coefficient the evaluator is concerned with the relative effectiveness of the test's prediction over what we would predict by merely guessing. Fortunately the degree to which a test improves upon chance can easily be determined by computing the coefficient of alienation, defined as the square root of one minus the validity coefficient squared or $\sqrt{1-r^2}$ Thus if a test has a predictive validity coefficient of .86, we are 50% more accurate in predicting the criterion using the test than we would be without it.

Another point to consider in evaluating empirical validity is the degree to which the test has been **cross validated**. Cross validation refers to establishing test validity on the same criterion for a group of persons different from those on which the original relationship between test and criterion was established. Cross validation can help the evaluator determine if the test is likely to be valid for program participants.

In evaluating **concurrent validity** the evaluator needs to consider whether or not a restriction of range has taken place. Briefly, restriction of range can occur when the criterion group consists of only those persons who are "successful" on the criterion behaviors. Persons who failed, or for reasons of their own left the criterion situation, are deleted from the group. This means that the measure being validated will be based on only "successful" cases. We would, of course, want to know the comparative test scores of criterion group failures and successes.

Unit Three

Module Five

Review

1. Validity may be defined as ____
____.

2. A measurement device must be ____________________ before it can be ____________
but it can be __________________________ and not be _________________________.

3. The two types of empirical validity are ________________________________ and
____.

4. Match the following:

    1. face validity
    2. predictive validity
    3. concurrent validity
    4. construct validity
    5. content validity

    a. usually established for achievement tests

    b. based on test appearance

    c. establishes a relationship between a group's test and ultimate outcome

    d. largely used to validate theoretical phenomena

    e. no followup on individuals is needed to establish the relationship between test and criterion

5. A predictive test that correlates .70 with the criterion adds _____% to our predictive ability.

References

Kunce, J., Miller, D., & Cope, C. Macro data analysis and rehabilitation program evaluation. **Rehabilitation Counseling Bulletin**, 1974, **17**, 132-140.

Unit Three

Module Six

Sources of Data in Vocational Rehabilitation

In reality, there are only two distinct approaches to evaluating a program. The first approach is purely subjective whereby an authority makes a judgment or issues an opinion as to the worth of a program. The second is more objective, based on scientific measurement, and preferred because it generates quantifiable data open to public inspection and subsequent confirmation or refutation. In this module four kinds of measures — existing records, questionnaires, rating scales and psychological tests — are considered as they relate to program evaluation in vocational rehabilitation.

Selection of measurement devices or "sources of data" should follow the principles discussed earlier in this unit. Program evaluation generally should not be based on a measurement instrument merely because the instrument is available. An exception may occur when a program evaluator is asked to evaluate a program after the program is in existence. In this case, the evaluator would, of course, be limited to those data which are routinely collected. In the selection of measurement instruments the evaluator is concerned with the purpose of the evaluation (e.g., assessing agency effectiveness and accrued client benefit), the type of change expected (absolute or relative), and the unit of analysis (agencies or individuals). Ideally, program objectives can be specified in behavioral terms. If so, program outcomes can be defined operationally. For example, a person is considered trained upon completion of a certain number of tasks or an individual is considered employed when certain preset criteria are met. Unfortunately, some of the most relevant program objectives are the most complex and, behaviorally speaking, the least precise. It is here that the evaluator becomes concerned with the selection of the evaluation instrument to meet the purposes of the evaluation. In selecting an evaluation instrument the evaluator needs to be concerned with reliability and validity, cost, time needed to complete the instrument and special modifications required if handicapped persons are to complete the instrument.

Existing records are forms used to categorize bits of data routinely collected by agencies and facilities. Generally, existing records are required for governmental reporting and are used in policy planning and programming. In rehabilitation the best known of these forms is the Statistical Reporting Form R-300. The R-300 provides a record of any client throughout the rehabilitation process. In that the R-300 has over 60 client related variables, and program related variables such as case expenditures, and is readily available and comparable across and within state agencies, it is a valuable source of data for the program evaluator.

The R-300 is a factual record. However, because it is completed and transmitted by people, the R-300 is subject to error. Formal reliability studies on the R-300 have not been completed. However, one study documented that there can be problems in relying on the R-300 as a data source. A regional program evaluation study group (Rubin, 1975) sought to examine agency operational efficiency by analyzing selected data from the R-300. This study was concerned with agency utilization of available resources and examined the percentage of clients receiving third party funds. The findings revealed that 83% of the third party services actually received by clients were not reported on the corresponding R-300 forms. Furthermore, the study group could only document 39 of 125 cases reporting no-cost services. Their conclusion regarding these R-300 data elements was that inaccurate reporting precluded the use of the R-300 in this type of program evaluation study.

Even factual recording forms such as the R-300 are subject to measurement errors. The program evaluator should make periodic checks to make sure that the data on the R-300 is accurate.

The most prevalent use of the R-300 is in summary statistical reporting. The annual publication **State Vocational Rehabilitation Agency Program Data** is a Federal document utilizing R-300 data. As such, the R-300 provides descriptive data regarding different agency operations and outcomes. On a national level R-300 data can be utilized to make comparisons among state programs and to draw inferences regarding program services. On the other hand, the R-300 when used on an individual level can be beneficial in isolating client or program characteristics differentiating successful from unsuccessful rehabilitants. When using the R-300 in this manner, the evaluator seeks to obtain R-300's on selected clients.

Other existing records can aid the program evaluator. Most training facilities routinely collect information on their clients. To the degree that such information is standardized, it can be used to supplement other program measures. Federal, state, and community organizations often collect program-related materials and provide another rich source of data. In Arkansas, for example, a quasi-public organization has spent considerable time and effort in surveying the spinal cord injured in the state. For an evaluator faced with evaluating rehabilitation services to the spinal cord injured, such a data source can be invaluable in such activities as estimating vocational rehabilitation penetration rates. Linkages with other agencies (e.g., Social Security, Internal Revenue), given the proper safeguards for confidentiality, can further add to the evaluator's data collection repertoire.

Existing records and the R-300 in particular are really structured reporting forms used to record existing measurement. **Questionnaires** are first order information collecting instruments. Questionnaires used via the survey technique are one of the most popular ways to collect client information. Questionnaires have great appeal since they represent a low cost method for collecting data on a relatively large sample. The major drawback of questionnaires is their vulnerability to response bias. Persons responding may answer any one question in terms of its social desirability; that is answering a question as one thinks the majority of people would answer it.

Quesionnaires have an intuitive appeal in supplementing other forms of evaluation data and they are most valuable in assessing the client's point of view. As an after service follow-up device, questionnaires can be used to assess program effectiveness (e.g., client satisfaction with services). Indeed, questionnaires have been very popular in vocational rehabilitation and have been used extensively in client follow-up studies and in the evaluation of inservice training programs.

Because of their intuitive appeal and apparent ease of construction, questionnaires continue to be popular. There are, however, some problems in using questionnaires in evaluation research. Reliability and validity are often not estimated for questionnaires, mainly because they tend to be used in one-time-only studies. Such one-time-only use also limits inter and even intra agency comparability. In constructing questionnaires evaluators should give attention to several factors. Careful thought should be given to the form of the questions; close-ended questions produce categorized data at the expense of response-richness obtained from open-ended questions. Open-ended questions are, however, very difficult to codify and summarize. A compromise is to use both types of questions. Attention should be given to reading level (sixth grade is about right), questionnaire length (10 to 15 minutes completion time), clarity of directions, response modification required for certain handicapped groups, and assurances of confidentiality. When used as a measure of the client's point of view and as a correlate of other evaluation measures, questionnaires provide a worthwhile addition to the evaluation effort.

Ratings and **rating scales** remain a popular and important measurement device in psychology, industry, and rehabilitation. Indeed, one of the earliest psychological measurement instruments was the man-to-man rating procedure developed by the Army during World War I. Griffiths (1970) defined rating scales as "essentially a means of recording (in quantified form) the impressions which an individual makes on others in terms of specified aspects of his behavior"(p.109).

Basically, there are two procedures used in ratings: one is a ranking relative to other persons on the same behavior; the other is scaling a person's behavior with respect to predetermined levels of the behavior. The scaling procedure generally scores a person's behavior somewhere along a continuum of successive intervals (e.g., from bad to average to good). The degree to which rating scales are behaviorally anchored, (i.e., use discrete descriptions of observable behaviors to define the intervals) and the degree to which raters have the opportunity to observe the behaviors affect both scale reliability and validity. As the rating scales depart from observable behaviors and demand the raters to infer abstract traits or characteristics in those being rated, they tend to become less reliable. Reliability, in this case, means the degree to which two or more observers agree in their ratings. Of course, the more abstract the rated behavior the less likely are two persons to agree to its occurrence.

Rating scales are particularly vulnerable to errors stemming from human characteristics most people exhibit when asked to judge others. For example, most people, in rating another person, tend to be overly generous in their judgments. This "error of leniency" acts to reduce the discrimination power of the scale. Halo effects, or rating someone based on overall impression, tend to displace all ratings toward the average. These and other rater errors such as the error of central tendency, or avoidance of scale extremes, subject ratings to a systematic bias which limits the usefulness of the device.

Rating scales are particularly susceptible to measurement error. Training raters to be aware of common errors can reduce this effect.

In vocational rehabilitation rating scales are a popular measurement device. While they can be constructed to fit practically any situation (i.e., categorizing obtained questionnaire scores, or level of interpersonal skill), rating scales are most prevalent in facilities and training centers where they serve as a measure of program effectiveness. Bitter and Bolanovich (1970) have developed a rating scale, the **Work Adjustment Rating Form**, that they say "can predict job adjustment with fairly high interrater agreement" (p.616). And the Arkansas Rehabilitation Service is experimenting with a rating scale, the **Facility Outcome Measure** (FOM), to supplement traditional outcome measures. The FOM rating scale consists of 22 items divided into three areas: vocational maturity, vocational functioning, and interpersonal maturity. Each item is followed by a five interval scale which carries a descriptor. For example, following the item "vocational skill level," interval one is "unskilled"; interval five "performs skill independently." A rater, presumably one who knows the client, rates the client on each item by checking one of the intervals.

Preliminary studies using the **Facility Outcome Measure** illustrate some of the problems in developing rating scales. Cooper (1976) sought to estimate, using interrater agreement, the reliability of this measure. He found that the three subscales all had reliability estimates of .49. These low estimates did improve over time ranging from .79 to .80 on the third paired ratings. The improvement in rating reliability may have been due to increased rater familiarity with the clients or to the previously mentioned rater errors. It should be noted that rater errors, if they are indeed systematic, will increase the estimate of interrater reliability (defined as interrater agreement) at the expense of scale discrimination. Furthermore, a study by Roessler, Cook, and Lillard (1976) found a high intercorrelation between the three subscales, suggesting that the subscales were not independent of each other. In reality, this particular scale is experimental; yet the above findings reveal the difficulty in the apparently simple task of developing a viable rating scale.

In the development of a rating scale, several procedures should be followed. First, select scale labels which are short yet clearly understandable, and avoid labels which make use of such words as excellent, reasonably, and occasionally. These words have different meanings to different raters. The more specific the cues the better. Second, use between five and eleven rating intervals for any one item. Third, select raters who are familiar with those being rated. Finally, spend some time training the raters, discussing the types of rating errors and the effect such errors have on the ratings. It is advisable to have the raters practice and discuss their results in order to reach agreement on what the various labels mean. Given careful scale development and rater training, rating scales can provide a unique source of data for the program evaluation effort.

Psychological testing is synonomous with "mental measurement." In actuality, the different measuring devices discussed so far, with the exception of existing records, are, in fact, "tests." We will define tests and inventories as those individually completed paper and pencil instruments which are published nationally, have been extensively developed, analyzed, and standardized to the extent that interpretive norms are available. One of the main advantages of such tests is that they usually have verifiable reliability and validity estimates.

There are literally thousands of psychological tests covering the range of human cognitive, affective, and conatus functioning. Such tests range from those applied in a variety of settings with a variety of populations to those developed for a specific population in a specific setting. In vocational rehabilitation psychological testing is most often associated with diagnosis. However, psychological tests, in that they relate to inferred abstract constructs such as self-concept or psychological adjustment, are the only measurement devices available when such constructs are important in the evaluation scheme. Unfortunately, relatively few of the thousands of psychological tests available were developed specifically with rehabilitation concerns in mind.

An example of a psychological inventory developed for use in rehabilitation is the **Human Service Scale** (Reagles & Butler, 1976). The scale consists of 80 multiple choice items scored on seven psychological needs thought to relate to Maslow's hierarchy. The Human Service Scale (HSS) is a relatively new instrument but is promising since it may provide a data source not previously available to rehabilitation evaluators. The HSS could be used as a measure of client change as a result of rehabilitation services in areas not tapped by reporting forms (e.g., the R-300). The HSS could also help differentiate the most severely disabled from less severely disabled, not on medical or functional grounds, but in terms of actual service needs. If program goals include improved client "psychological adjustment," the HSS could be a valuable criterion measure for measuring program **efficiency** (services expended in meeting scale-defined client needs), **effectiveness** (documenting sustention of benefits through postservice attainment of needs), and **quality** (relative amount of needs met).

The **Attitudes Toward Disabled Persons Scale** (Yuker, Block, & Younng, 1966) was developed as a measure of the attitudes of nondisabled persons toward the disabled. The test is essentially a self-rating scale by which the respondent lists agreement-disagreement to each of fifty items. Currently, because the relationship between client attitude and actual behavior is not well understood, attitude change is not seen as a particularly valid measure of rehabilitation service effectiveness. However, as a correlate of client success (or failure) or as a measure of employer or community attitudes affecting client outcomes, this measure can provide valuable information not attainable elsewhere. For example, type of client attitude might be an important variable in caseload management. Use of the ATDP by Krauft, Rubin, Cook, and Bozarth (1976) illustrates the use of this measure in a process or formative type of evaluation. Using the ATDP Krauft, et al. (1976), found that counselors who held more positive attitudes toward the disabled experienced greater success than less positive counselors when working with clients whose disabilities were perceived most different from the able-bodied. The implication is that counselor attitude toward the disabled does relate to successful counseling.

Psychological tests can add to the evaluation scheme in that they are estimates of a person's inner state (e.g., personality and attitudes). Other advantages include, they may be used pre and post service to assess degree of gain, test scoring is objective leaving little room for subjective judgments, and standardized norms which facilitate interpretation of scores. Disadvantages include limitations in meeting particular program needs, prohibitive administration time, cost, and by their very nature they are susceptible to subject reactivity (see Unit Four, Module One).

All measurement instruments have advantages and disadvantages as related to the goals and objectives of program evaluation. Ideally the program evaluator would utilize those measures most adaptable to assessing program effectiveness, efficiency and quality.

Unit Three

Module Six

Review

1. In the selection of measurement devices the evaluator needs to be concerned with the ________________, the ______________________ and the ______________________________.

2. Four kinds of measurement devices are _______________________, _________________________, _____________________________________,and _______________________________________.

3. What are some important considerations in formulating questionnaires?

4. In ratings, the two measurement procedures used are
   a. ____
   b. ____

5. Rating scales most often use the scaling procedure of ____.

6. Rating scales are vulnerable to the measurement errors of
   a. ____
   b. ____
   c. ____

7. Discuss how the four types of measures (existing records, questionnaires, rating scales and psychological tests) might be utilized in a comprehensive assessment of rehabilitation agency effectiveness and efficiency.

References

Bitter, J., & Bolanovich, D. WARF: A scale for measuring job-readiness behaviors. **American Journal of Mental Deficiency,** 1970, **74,** 616-621.

Cooper, P. The assessment of rehabilitation facility service in Arkansas. **Research Report.** Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1976.

Griffiths, R. Personality assessment. In P. Mittler (Ed.), **The psychological assessment of mental and physical handicaps.** London: Tavistock Publications, 1970, p. 83-131.

Krauft, C., Rubin, S., Cook, D., & Bozarth, J. Counselor attitude toward disabled persons and client program completion: A pilot study. **Journal of Applied Rehabilitation Counseling,** 1976, **7,** 50-54.

Reagles, K., & Butler, A. The Human Service Scale: A new measure for evaluation. **Journal of Rehabilitation,** 1976, **42,** 34-39.

Roessler, R., Cook, D., & Lillard, D. Effect of systematic group counseling on work adjustment clients. **Journal of Counseling Psychology,** 1977, **24,** 313-317.

Rubin, S. Studies on the evaluation of state vocational rehabilitation agency programs. Arkansas Rehabilitation Research and Training Center, University of Arkansas, Fayetteville, November, 1975.

Yuker, H., Block, H., & Younng, J. **The measurement of attitudes towards disabled persons.** Albertson, NY: Human Resources Center, 1974.

97

UNIT IV

Data Analysis and Research Design for Program Evaluators

98

TABLE OF CONTENTS

Module Page

Traditional Evaluation Research Design 109
Objective: To introduce the traditional concepts and discuss techniques for evaluative research design.

Alternative Evaluation Designs 117
Objective: To discuss the utility and limitations of alternative designs such as ex post facto studies and cost benefit analysis.

Introduction to Data Analysis Through Statistical Hypothesis Testing 123
Objective: To introduce statistical hypothesis testing.

Representative Techniques for Data Analysis 129
Objective: To discuss various types of statistical inference problems a program evaluator may encounter and suggest potential statistical techniques which may be applicable.

Confidence Interval on a Mean 135
Objective : To introduce the concept of confidence interval and illustrate some computational aspects.

Testing for Differences in Means 139
Objective: To introduce various techniques for testing differences in means.

The Sign Test 147
Objective: To illustrate the utility and computation of the sign test.

Chi-square Tests 151
Objective: To illustrate the computation and interpretation of Chi-square tests in contingency tables.

The Language of Data Processers 159
Objective: To introduce some of the basic vocabulary used in data processing.

Considerations in Presentation 163
Objective: To emphasize the importance of adequate presentation of evaluative research results.

Evaluation Research: Annotated Bibliography 171

99

Unit Four

Module One

Traditional Evaluation Research Design

Program evaluators attempt to answer whether or not a program or project has met its stated goals. In modern day research jargon the evaluator wants to determine **casuality**, i.e., whether or not the program caused the observed and hopefully intended outcomes. Searching for casuality follows from "if-then" statements: If the program realizes its objectives, then more disabled people become employed. Causality considers "why did it work?" type of questions. "What happened?" questions seek to describe program outcomes while "Why did it happen?" questions go beyond description of outcome to specification of what caused the outcomes. This latter determination is much more important than the former. Desired program changes such as an increased client employability may be observed and yet be unrelated to the program's interventions. If desired outcomes occur without being linked to a program's interventions, the program would have a reduced chance of survival. Conversely, if a program can be shown to have caused the desired effect, accountability will be enhanced.

Hartley, (1976),an authority on experimental design, noted that scientific research generally proceeds by formulating hypotheses which are then verified either directly or through their consequences. The verification process includes the collection of observations and the **experimental design** consists of the plan by which the observations are to be collected.

Behavioral sciences have evolved to the point where there are generally accepted designs, each having advantages and disadvantages, regarding the inference of causality. Various experimental designs differ in comprehensiveness, practicality and power to discredit rival explanations of causality.

These different types of research designs are generally divided into three categories: pre-experimental, quasi-experimental, and true experimental. Campbell and Stanley (1963) in their classic book on design list three pre-experimental, nine quasi-experimental, and four classes of true experimental designs. The degree to which the analysis of date obtained through one of the designs answers the question "why did it work?" depends upon the extent to which the research design is able to rule out plausible alternative explanations, i.e., effects caused by something extraneous and unconnected with the program.

The **sine qua non** of evaluation research is that the data of a program evaluation study are valid. Campbell and Stanley (1963) made a distinction between the internal and external validity of a research design. Internal validity refers to the degree to which design eliminates extraneous explanations to the causality question. External validity refers to the generalizability of the findings. According to Campbell and Stanley there are eight classes of extraneous variables which can affect internal validity and four classes which might affect external validity. Threats to internal validity include:

history,events and happenings that occur between the beginning and end of a program, for example an economic recession;

maturation, growth and changes within the individuals exposed to the program which occur independent of the program;

testing, the effect that completion of a test has on a second completion of the same test; and,

mortality, differential loss of persons from comparative programs.

Threats to external validity include:

reactive effects of testing, where pretesting may increase or decrease an individual's sensitivity to program services; and,

reactive effects of experimental arrangements, where the arrangement of program sevices precludes drawing generalizations to persons not exposed to the services.

The idea is for the evaluator, with given situational circumstances, to choose the best design to limit the above threats.

Quasi-experimental Design

Perhaps the most common of the quasi-experimental designs is the **one group pretest-posttest design.** This design is illustrated in Figure 1.

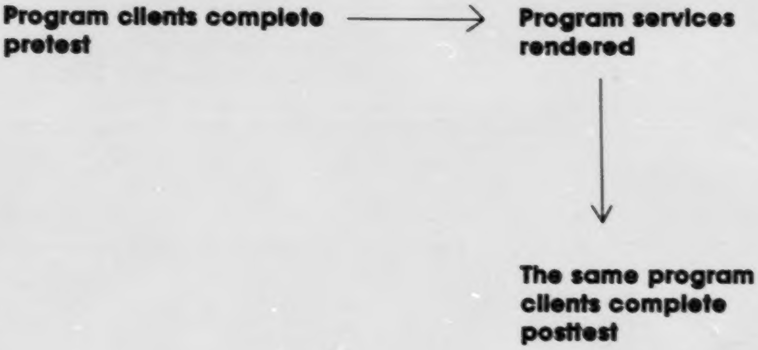

Program clients complete pretest
Program services rendered
The same program clients complete posttest

Figure 1. One Group Pretest-Posttest Design

An example of the use of this design is illustrated below.

> Rehabilitation counselors in four states are selected by their supervisors to attend a two day training session on the Individualized Written Rehabilitation Plan (IWRP). The trainers are concerned that their training is effective,so they have the trainees complete a test of knowledge on the IWRP before and after training. Assuming that the test is valid and reliable the trainers compute post minus pre difference scores for each participant and average these difference scores to obtain a mean difference score. Using an appropriate statistical test, they discover that test scores after training are significantly higher than test scores before training. They conclude that the training was effective.

Unfortunately the one group pretest-posttest design jeopardizes internal validity. There are possible alternative explanations, other than training, to account for the gains made by the group of trainees. **History,** or changing events not associated with training may have affected the outcomes. For example, a law requiring use of the IWRP may have come into effect between pre and posttesting. History is a plausible rival explanation for the increase in test scores. On the other hand, trainee maturation, or changes in the trainees which occurred with the mere passage of time may have affected the observed differences. Finally the effect of **testing,** the knowledge the trainees gained by completing the pretest may have influenced the observed gain. While the one group pretest-posttest design is obviously better than subjective judgements, there are often enough un-controlled rival explanations to render any inferences about training effects suspect.

Another quasi-experimental design is the **nonequivalent control group** design. In this case the one group pretest-posttest design is improved by adding a control group or a group of similar trainees who do not receive training (see Figure 2).

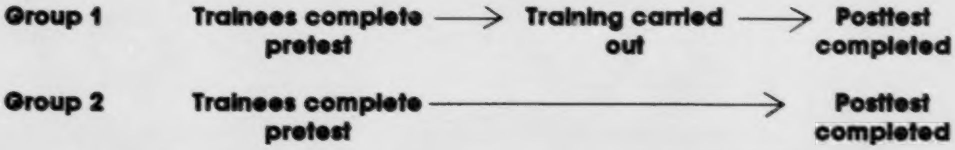


Figure 2. Nonequivalent control group design

Assuming (1) that persons in the training group and non training (control) group are similar and (2) that those in the control group did not show the gains of IWRP knowledge which those in the training group did, then history, maturation, and testing can be ruled out as plausible alternative explanations for the observed gains. However, the nonequivalent control group design does not control for another threat to internal validity--group selection bias. It is possible that persons who received training were more knowledgeable about the IWRP to begin with and that this head start interacted with individual maturation rate to produce exaggerated gains in knowledge.

The final quasi-experimental design to be considered is the **interrupted time-series design.** Basically the time series design utilizes a series of measurements or tests interrupted by the intervening treatment (i.e., program services, training, etc.) The time series design (see Figure 3) controls for maturation, testing, and the interaction of selection and maturation. In effect when an intervention procedure is analyzed via this design the subjects serve as their own control group.

Training group Test→Test→Test/Training/→Test→Test→Test
1 2 3 4 5 6

Figure 3.
Illustration of the Interrupted
Time Series Design

Figure 4 illustrates different interpretations which can be drawn when evaluating a program via the interrupted time series design. Attributing test score changes to the program intervention is more valid for groups A and C, both groups had stable pre-program intervention test scores followed by post-program intervention test score increases. On the other hand, group B had a constant straight line increase across all test scores and group D had a similar increase both before and after the intervention, effects not likely to be associated with the intervention.

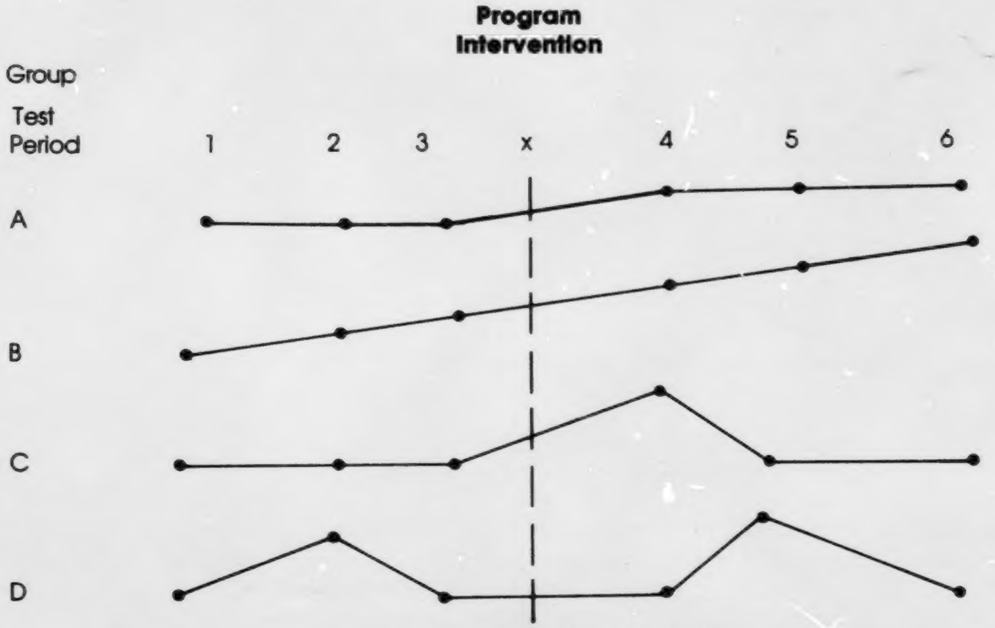


Figure 4. Alternative results for different groups using interrupted time series design. X is the point of program intervention

True Experimental Designs

The most common example of a true experimental design is the pretest-posttest control group design (Figure 5). The way in which this design differs from the nonequivalent control group design (Figure 2) is that subjects are assigned at **random** to both the training (experimental) group and the non-training (control) group. The key to the design is the random assignment of persons to the experimental and control groups. Random assignment equalizes both groups in terms of such things as prior knowledge, level of achievement and so on. Theoretically this design controls for all threats to internal validity. Because random assignment has equalized the groups, the effects of, for example, history and maturation should be the same for both groups. If in figure 5, Posttest B is significantly greater than Posttest D, then the evaluator can say, with confidence, that training was the most likely cause of the difference.

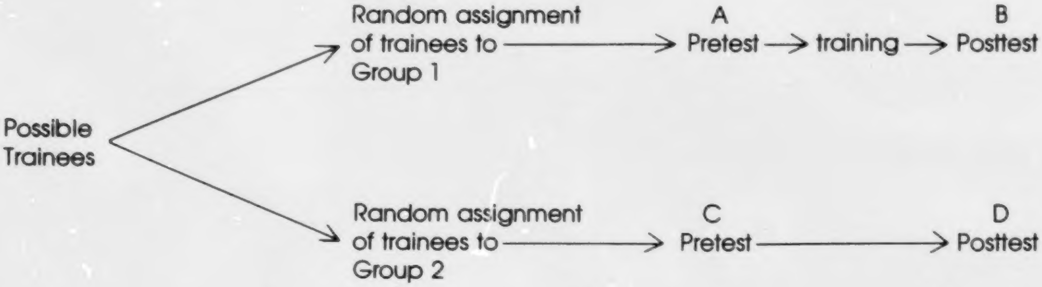

Possible Trainees
Random assignment of trainees to Group 1
A
Pretest
training
B
Posttest
Random assignment of trainees to Group 2
C
Pretest
D
Posttest

Figure 5. Pretest-Posttest Control Group Design

There is one drawback to this design. That is the process of measurement itself may affect the trainees subsequent performance. In other words, changes in observed scores from pre to posttest may not be due to training but to the effects of testing (i.e., trainees learning how to respond "correctly" as a result of taking the pretest). This drawback can affect the generalizability of the results, and may limit whether or not we think the training would be just as effective for trainees not exposed to a pretest warm-up. Generally, this is not a major concern in evaluation research especially in large scale evaluations. The rigorous evaluator can make statistical corrections, for example, using an individual's pretest score in conjunction with the posttest score to determine the relative degree of measurement influence.

One way in which measurement reactivity can be deleted as an experimental error is to employ the **posttest-only control group design.** This design is identical to Figure 5, except that no pretests are given. This design, given random assignment to treatment and control groups can be efficient. It is probably used less than other designs for a number of reasons, not the least of which is the evaluator's desire to determine if randomization was effective, (i.e., the groups did not differ in terms of the pretest). If however, the principle of randomization is carefully followed both groups will of course be equal. Giving up pretest information can, nevertheless, limit the evaluation. This is especially true when we want to know how **much** the program impacted on the participant. This is of course very difficult without preprogram measures.

Beware of Extraneous Stimuli

Hartley (1976) pointed out that comparative evaluations involve the measurement and comparisons of "responses" from identical experimental units exposed to specific "treatments" applied by the evaluator. Further, the elimination of all extraneous stimuli which could potentially influence the response is essential. The elimination of such stimuli is referred to as experimental control.

There is no dearth of examples which illustrate the potentially disastrous effects of failing to control for extraneous stimuli or latent characteristics. Hartley (1976) illustrates this point by showing a strong linear relationship between the live birth rate among married women and the total number of storks observed in Swedish sanctuaries. In this case the extraneous stimulus was an increase in living standards in Sweden which raised both the birth rate and the number of sanctuaries the Swedes could afford. Assuming that the number of storks is the treatment and the live birth rate the response, the extraneous stimuli (living standard) affected **both** the treatment and the response.

Although this example may seem ridiculous, it is quite possible for a program evaluator to fall into the same trap when the relationships are more subtle. Consider for example, the time Charlie Numbercruncher compared two district offices in a vocational rehabilitation agency with respect to the average yearly earnings of their successfully rehabilitated clients. One office was in a rural area where most employment opportunities were for farm labor while the other office was located in the state capital where average wages were significantly higher. Charlie found the state capital offices had higher paid rehabilitants and concluded at first that they were therefore doing a superior job in placement. He subsequently realized that this conclusion was erroneous since he had ignored the latent characteristic of environment.

In another instance, Charlie was approached by an instructor at a vocational rehabilitation facility wishing to evaluate the effectiveness of a reward system for improving student ability to remain at a work station. Charlie recommended a pretest-posttest type design with the treatment period running for four weeks. At the end of the period the group did show marked improvement in their ability to remain at a work station. It was not until a second instructor at the facility asked Charlie to evaluate his ongoing Work Adjustment Training program that Charlie realized he had ignored a very important extraneous stimuli — a substantial portion of the clients in Charlie's sample were simultaneously enrolled in work adjustment training.

Unit Four

Module One

Review

1. Traditional research designs seek to limit threats to ______ and ______ validity in order to determine ______.

2. What is the major difference between quasi-experimental and true experimental design?

3. Threats to internal validity become plausible ______ ______ for observed effects.

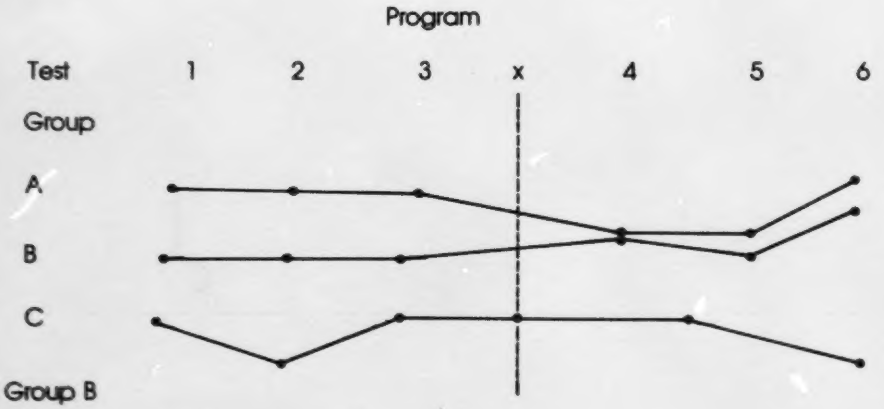
4. Match the following:

   1. An agency, independent of an innovative service program, cuts back its staff
   2. Completing a test gives knowledge for the correct way to complete a subsequent test
   3. The control group loses 25 subjects, the experimental group loses 2 subjects
   4. There is a lapse of two years between pre and post-testing

   a. Effect of testing
   b. Differential Mortality
   c. Effects of History
   d. Maturation

5. True experimental designs control all threats to ______ validity.

6. Which one of the following programs was the most successful?



Answer: Group B

References

Campbell, D. T., & Stanley, J.C. **Experimental and Quasi Experimental Designs for Research.** Chicago,Ill.: Rand McNally, 1963.

Hartley, H. O. The design of scientific components. **Sigma XI Lecture.** Fayetteville, Arkansas: University of Arkansas, 1976.

Unit Four

Module Two

Alternative Evaluation Designs

While true experimental and quasi-exerimental designs are the best way for the evaluator to study the effectiveness of a program there are times when those designs are not applicable. Such is often the case in large scale evaluations of entire national programs; when random assignment of persons to experimental and control groups or even the formation of nonequivalent groups is impossible; when the evaluator arrives late on the scene, the evaluation is underway and traditional pretesting is impossible; or when the evaluator is asked to set up a program monitoring system focusing on a formative evaluation (ongoing program feedback) rather than a summative (terminal evaluation of program worth).

Ex Post Facto Evaluations

In vocational rehabilitation relatively few evaluations or traditional research studies for that matter, follow true experimental designs. Rather most evaluations are **ex post facto** (literally "after the fact") studies. The drawback to such evaluations is that one only knows how persons function after program services. It is difficult to determine any gains made by program clients and almost impossible to attribute post program functioning to program services received.

Ex post facto evaluations are one-time only studies and while they are vulnerable to almost all the threats to internal and external validity they are not without value. For example, sampling program participants' satisfaction with the program can provide valuable program feedback. By forming groups of program participants based on available preprogram data such as age and analyzing which groups have favorable or unfavorable program outcomes, the evaluator can often make valuable recommendations for improved service provision **vis a vis** different types of persons. Such studies are said to be descriptive and are concerned with documenting results. Note however that descriptive studies cannot show that program effects were a direct result of program sevices.

With the increased emphasis on accountability, Federally funded service programs and projects are mandated to include an evaluation component. As is too often the case the evaluator is called in after the program is well under way, or even completed. In the latter case it is sometimes difficult to obtain relevant program outcome data. Thus the evaluator attends to **bridging variables** and an evaluation of **program processes.** Bridging variables are those things thought to relate to ultimate outcomes. For example, in vocational rehabilitation employment is the ultimate outcome. Certain psychological outcomes such as improved self concept and increases in coping ability are thought to relate to employability. Thus psychological changes and such intermediary outcomes as graduation from vocational training are considered bridging variables and in the absence of measures of ultimate outcome become the criteria of choice.

On the other hand in some human service programs the evaluation emphasis is not on program outcome, but on the effectiveness of program service provision. The evaluation in this case is process oriented. The argument for process evaluations is that services are legally and ethically mandated, and since there isn't a question of whether or not services will be continued it behooves the evaluator to determine the most effective way to deliver the services. The counter argument is of course, why have programs if they can't be shown to impact on those served by the program?

In assessing the effectiveness of program processes the evaluator is concerned with measuring the scope of services delivered compared to the expected extent of services to have been delivered, the quantity of services delivered, and the relative speed at which services were delivered. There are at least two ways an evaluator can assess program processes. If the

107

evaluator is called in when the program is underway it may be possible to collect "pre" program service data on persons just entering the program, and "post" data on persons midway through the program and on persons who have completed the program. Assuming there were no obvious changes in program recruitment policies,it may be possible to relate program services received to participant progress. Alternatively, in programs that routinely collect certain data, the evaluator can make use of that information. For example, in vocational rehabilitation a standardized reporting form, the R-300, is completed for each person served. Utilizing the R-300 it is possible to assess such things as client time in status and case movement.

Comparative Evaluation

One of the real advantages of true experimental designs is that they produce data which have universal understanding. For comparison evaluations the true experimental designs are by far the best. Lacking the resources or support for experimental evaluations the evaluator can make comparisons between programs, and between subparts of a program, by using prespecified performance standards.

For example, the Rehabilitation Services Administration is currently evolving a set of "standards" or performance criteria that will eventually carry the force of regulations. These standards in conjunction with the annual service goals of each state permit a nationwide summary of agency performance in the areas of sustention of vocational rehabilitation benefits, client satisfaction, suitability of placement, and adequacy of services. Given standard reporting, across state comparisons on performance among the subunits of the vocational rehabilitation program can sometimes be made.

The West Virginia Rehabilitation Research and Training Center (Moriarty, 1977) has developed a Profile Analysis Technique (PAT) which can be used as an inter or intra agency comparison measure. Basically PAT makes use of data transformations (see Unit Three) to plot such things as time in status, case closures, number of rehabilitants per counselor, and number of severely disabled rehabilitated. In and of themselves, the above data elements lack meaning. In the PAT procedure the raw data are transformed to normed scores that have a common reference point. Using the above procedure it sometimes becomes possible to compare different agencies, or counselors within a single agency using well defined reference points.

However, PAT is limited to rehabilitation information routinely collected by an agency. The technique is less effective in plotting scores for various intangibles. Walker (1972) has combined goal setting and feedback into a procedure that has promise for program monitoring. Basically, the system involves (1) developing expected outcomes regarding program goals, (2) ordering the goals into priorities, (3) attaching weights to the goal priorities, (4) measuring outcomes in terms of the assigned weights and (5) providing feedback to program personnel as to how they have done in relation to program priorities. One advantage to this approach is that program planners are forced to formulate relevant and practical goals. The agreed upon weighting system serves as a reference for determining the relative effectiveness of program subcomponents.

Cost-Benefit Analysis

Cost-benefit analysis can be defined as follows "... an analytical study designed to assist a decision maker in identifying a preferred choice among possible alternatives ... an attempt to minimize dollar cost, subject to some mission requirement (which may not be measurable in dollar terms) or conversely to maximize some physical measure of output subject to a budget constraint" (Bennett & Weisinger, 1974, p. 13). Cost-benefit analyses are often used in program planning to estimate the forecasted cost of a planned program relative to the estimated dollar return on program benefits. Weiss (1972) succinctly stated this procedure.

> In essence, the cost-benefit analyst attempts to identify the benefits of a program, both tangible and intangible; he looks at the costs of conducting the program, the direct and indirect; then he tries to put them into a common unit of measurement-dollars. The ratio of benefits to cost is an indication of the return that society is getting from its investment in the program. (Weiss, 1972, p. 85).

As might be expected, one of the problems in cost-benefit analysis is trying to quantify in dollars such program participant benefits as increased well being or reduced psychological distress. Vocational rehabilitation traditionally has made use of cost-benefit analysis on a national level. For example, showing the benefit in tax income when a person is rehabilitated. Probably the real value to cost-benefit analyses is that they produce figures with real meaning to lay persons and, if favorable, carry considerable clout in the political decision making process.

Unit Four

Module Two

Review

1. Ex post facto studies are ______________________ in nature, no cause and effect conclusions can be drawn.

2. Ex post facto studies can provide a program __________ function.

3. When the evaluator is unable to formulate preprogram measures he must turn to __________ variables and measures of ____.

4. The argument for conducting efficiency evaluations in mandated service programs is: ____.

5. How might the PAT technique be used to compare two rehabilitation programs? ____

6. Costs-benefit analysis might be defined as: ____

References

Bennett, E.C., & Weisinger, M. **Program Evaluation: A Resource Handbook for Vocational Rehabilitation.** Springfield, Ill.: Charles C. Thomas, 1974.

Moriarty, J.B. Profile analysis technique. In Muthard, J.E., (Ed). **Counselor and Program Evaluation Methods in Vocational Rehabilitation.** Gainesville, Fl.: Rehabilitation Research Institute, 1977, 1-13.

Walker, R.H. The ninth panacea: program evaluation. **Evaluation**, 1972, **1**, 10-19.

Weiss, C.H. **Evaluation Research.** Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Unit Four

Module Three

Introduction to Data Analysis Through Statistical Hypothesis Testing

Assume that a program evaluation specialist has chosen a design, selected a sample and collected the data. The evaluator is now ready to analyze the raw data and interpret the results of the analysis. The process of data analysis traditionally goes beyond subjective judgement and relies on probability and inferential statements as mediums of expression. Statistical data analysis techniques are often useful in evaluation research. Effective utilization of these techniques requires the program evaluator to be familiar with the theory and rules for drawing inferences from the data. One such approach to inference is statistical hypothesis testing.

Statistics play a large role in our everyday life. We follow average points per game and average yards gained per game of our favorite football team. The price we pay for life insurance is based on the probability that persons our age will live only so long. In evaluation research statistical methods are the tools by which raw data can be classified, made manageable and hence interpretable. The use of statistical tools becomes especially important when the program evaluator wants to infer that some program **caused** some specified change. Unfortunately, the evaluator who follows the scientific method knows that making general statements from specific observations will always have some uncertainty and that even descriptions will contain errors or chance factors. Through appropriate use of statistical techniques, the evaluator can specify how much of the observation can be attributed to chance happenings.

The appropriate statistical technique depends upon the nature of the measurement device, the type of inference desired and the manner in which the data were collected. Thus, the level of measurement (for a review, see Unit Three), and the research design both place limits on which statistical techniques might be applicable. It is, for example, better to control experimental errors through improved research design than through post hoc statistical adjustments.

Generally, the appropriate statistical analysis is determined by the evaluator's hypothesis and is selected to document the truth or falsity of that hypothesis by isolating the error or chance variability which might explain the hypothesis. The research hypothesis is the key to the scientific method and to traditionally statistical analysis. Hypotheses usually state some theoretical relationship between two or more variables or populations. A hypothesis is never **proven** true or false, rather it is shown to be **probably** true by statistical analysis. Statements which can only be true (either I am human or I am not human) or statements which are always false (I am human and I am not human), are not testable hypotheses. A hypothesis must be capable of being true or false. Statistical hypothesis testing helps determine the most likely of these two states. The program evaluator interested in causality will formulate evaluation hypotheses from program goals in the form of "if . . ., then . . ." statements. Any hypothesis should of course be testable, relevant to the program goals and expressible in a quantified form.

In statistically testing a research or evaluation hypothesis it is common practice to state the hypothesis in a "turned around" form. For example, consider the evaluation hypothesis: "If the XYZ program is implemented, then program participants will exhibit significant gains in employability."Stated in the "turned around" or **null** form the hypothesis would read: "After program XYZ is implemented, there will be no change in the employability of program participants." In conjunction with any null hypothesis there must always be an **alternative** hypothesis which states what the evaluator will **assume** to be true if the null hypothesis is false. For the null hypothesis above, an appropriate alternative might read: "As a result of program XYZ, there will be an increase in employability of program participants." Ordinarily, that which the researcher wishes to demonstrate is stated in the alternative hypothesis. Thus, in any hypothesis testing situation the researcher must state a **null hypothesis** which he is willing to accept as true unless he receives overwhelming evidence to the contrary; in which case he may conclude that his **alternative** hypothesis is true. He assumes that exactly one of the two hypotheses must be true.

The research question above was concerned with a single population. Other hypotheses about single populations include **correlational hypotheses** or hypotheses about the relationship between two variables. For example, the null hypothesis:

H_0: The correlation between a client's educational level and his income at closure is zero might be tested against the alternative hypothesis,

H_1: The correlation between a client's educational level and his income at closure is greater than zero.

Another type of single population hypothesis involves statement of specified values for population parameters. For example, it may be desirable to test the null hypothesis,

H_0: The mean earnings at closure of rehabilitated clients is $116 per week (federal minimum wage),

against the alternative hypothesis,

H_1: The mean earnings at closure of rehabilitated clients is greater than $116 per week.

Other research questions often involve more than one population. This type of question is common in evaluation research and usually deals with questions concerning the relative value of alternative programs. For example, if two different programs, A and B, are designed to produce similar results, an appropriate null hypothesis might be:

H_0: Program A participants learned the same amount as Program B participants.

An alternative hypothesis might be:

H_1: Program A participants differed from Program B participants with respect to the amount learned.

Statistical Hypothesis Testing

The basis of traditional statistical inference is hypothesis testing. There are six basic steps to statistical inference based on a hypothesis test.

1. **Construct a general theory regarding some aspect of the program.**

Example: State A is about average with respect to its cost per closure among rehabilitated clients (26 closures).

2. **Refine this general theory into a hypothesis such that data may be collected to test the hypothesis.**

Example: The mean cost per closure, μ, among State A's rehabilitated clients is \$1000 (the national average). This can be rewritten as:

$H_0: \mu = \$1000$

This hypothesis is called the **null hypothesis**, and it will be **assumed** to be **true** unless the data collected indicate otherwise. The **alternative hypothesis**, H_1, is assumed to be true if in fact the null hypothesis is found to be false. For the null hypothesis above, the researcher has four choices for H_1:

1. $H_1 : \mu > \$1000$ one-sided alternatives
2. $H_1 : \mu < \$1000$
3. $H_1 : \mu \neq \$1000$ two-sided alternative
4. $H_1 : \mu = \mu_1$ (μ_1 is some specified value other than \$1000

To test the null hypothesis, μ will be estimated by the test statistic $\overline{x}$, the average cost of a random sample of cases.

Recall that null hypotheses regarding a single population usually state that some characteristic or parameter of the population has a specified value. Null hypotheses about more than one population usually state that the populations have a common value for one or more parameters.

3. Select a statistical test which can distinguish between the truth or falsity of H_0 and whose assumptions are satisfied by the data being collected.

4. Specify the **level of significance** (α), that is, the probability of rejecting H_0 when in fact it is true. Then, select the sample size n.

Any hypothesis test must result in one of four possible outcomes: accepting H_0 when it is true, rejecting H_0 when in fact it is true, accepting H_0 when it is in fact false and rejecting H_0 when it is false. Figure 1 depicts these four combinations. When the researcher rejects the null hypothesis when in fact it is true, he has committed a **Type I error.** Likewise, when he accepts the null hypothesis when it is false, he has committed a **Type II error.** Traditionally, researchers are more concerned about controlling the probability of a Type I error than a Type II error. The probability of a Type I error is denoted by α (see Figure 2) and is called the **level of significance** or size of the test. Consider, for example, pharmaceutical researchers. By emphasizing the importance of Type I error, they assert that it is better to make a Type II error in claiming a drug is ineffective when it really has some effect, than it is to make a Type I error and claim an ineffective drug is effective and release it for public consumption.

Figure 1
ERRORS IN HYPOTHESIS TESTING

State of	Decision	
Nature	Accept H_0	Reject H_0
H_0 true	No error	Type I error
H_0 false	Type II error	No error

Figure 2
PROBABILITY OF ERRORS

State of Nature	Accept H_0	Reject H_0
H_0 true	$1 - \alpha$	α
H_0 false	β	$1 - \beta$

Generally, α is specified in advance of the data collection. It is set by the evaluator and depends on how much the evaluator wants to guard against making a type I error. Convention in psychosocial research generally set the level at .05 (i.e., 5 times out of 100 the evaluator might reject the null hypothesis when in fact it is true).

The probability of a Type II error is denoted by β. Usually, after the probability of a Type I error (α) is specified, the only way to reduce the probability of a Type II error (β) is to increase the sample size. The quantity ($1 - \beta$) is called the **power** of the test and clearly increases as β is reduced by increasing the sample size. ($1 - \beta$) is the probability of rejecting H_0 when in fact it is false and hence is a function of the alternative hypothesis. Most tests are more powerful against **extreme** hypotheses which differ greatly from the null hypothesis. If the sample size is allowed to become arbitrarily large, both α and β may usually be specified. In most instances, small α and large power ($1 - \beta$) are desirable.

5. Identify the **critical region**, i.e., the set of outcomes leading to rejection of H_0. The critical region consists of those outcomes from the data collection which are so extreme that they must be incompatible with H_0 and therefore cast doubt upon the truth of the null hypothesis. That is, the critical region consists of all those outcomes which would be highly improbable if the null hypothesis were true.

The remaining set of outcomes is called the **acceptance region**. Boundary values between the two regions are called **critical values**. Based on the distribution of the test statistic when H_0 is true and consideration of the alternative hypothesis H_1, the critical region is chosen so that the probability is less than or equal to α and that the test statistic ($\overline{x}$, for example) falls in this region when H_0 is true.

Example: Let μ be the mean cost per closure for State A. Suppose \$1000 is the national average cost per closure.

H_0: $\mu = \$1000$

H_1: $\mu \neq \$1000$

Significance level = α

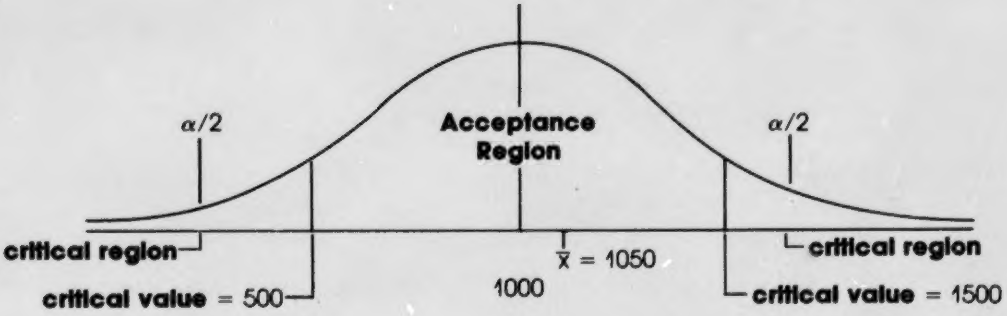
Observations: $x_1, x_2, x_3, \ldots x_n$

$$\text{Test statistic } \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \text{(say 1050)}$$

115

Figure 3

Distributions of $\overline{x}$ when H_0 is true



α/2
Acceptance Region
α/2
critical region
critical region
$\overline{x}$ = 1050
1000
critical value = 500
critical value = 1500

Since $\overline{x} = 1050$ is in the acceptance region we accept H_0 and conclude $\mu = 1000$. If x had been greater than 1500 or less than 500, the null hypothesis would have been rejected.

6. Decide whether to accept H_0 or reject H_0 in favor of H_1. If the test statistic falls in the critical region, then either

(a) H_0 is true and a very rare event has occurred;

or

(b) H_0 is false (in which case H_1 must be true by prior convention).

The probability that (a) is true has been specified as equal to α and is usually selected to be small. Hence, conclusion (b) is more likely. We say that a test statistic is **significant** if it falls in the critical region, i.e., if the probability of observing a value as extreme or more extreme is less than or equal to α.

Review

1. Write null and alternative hypotheses relevant to the following general theories:

Example: Male and female rehabilitation clients have the same average weight.

$H_0 : \mu_1 - \mu_2 = 0$	The difference in the mean weight for men and the mean weight for women is 0.
$H_1 : \mu_1 - \mu_2 > 0$	The mean weight for males is greater than the mean weight for females.

a) Severely disabled clients and non-severely disabled clients spend about the same length of time in referral status 00.

b) According to an alcohol treatment facility schedule, a client should remain at the facility for two weeks.

c) Vocational training costs for males and females tend to be about the same, for clients receiving this service.

2. Fill in the blanks.

a) The level of significance is the probability of ______ H_0 when in fact it is ______________.
b) The power of a test is the probability of _______ H_0 when in fact it is ______________.
c) By increasing the sample size for a fixed level of significance we can usually _______ the power of a test.
d) The power of a test is a function of the _______ hypothesis.
e) If the test statistic falls in the critical region, then H_0 is ______________________________.
f) If H_0 is disclosed to be rejected then the alternative hypothesis must be ______________.

Unit Four

Module Four

Representative Techniques for Data Analysis

After careful study of the theory of division of labor, Charlie, in a flash of brilliance, concluded that he should ask his faithful and loyal associate, P.J., to compile a list of the most common types of data analysis that program evaluators might encounter. In addition, for the sake of efficiency, he decided that P.J. might as well research possible approaches to each type of problem. With a great deal of self-assurance, he confronted P.J. with the fruits of his contemplation to which P.J. replied in a matter-of-fact manner: "Stick it in your ear, Charlie. It's not my job description!" Unwilling to be intimidated, however, Charlie suggested a compromise. "If I make the list of problems and possible approaches, will you proofread it and correct the spelling?" P.J. agreed and the resulting product appears below.

Problem — estimation of a mean. It is sometimes desirable to estimate the mean of a particular variable for a single population. For example, a program evaluation specialist may be charged with estimating the mean time in training for a specific population of clients. Some assessment of the precision of the estimate must also be made. A similar problem occurs when the program evaluator must estimate the proportion of a population possessing a specific characteristic. For example, it may be necessary to estimate the proportion of cases entering extended evaluation which do not have documented justification for the use of status 06 in the case folder. Again, an assessment of the precision of the estimate would be required.

The simplest way to estimate a mean is to compute a sample mean (usually denoted by $\overline{x}$). The most common way of assessing the precision of this estimate is to report the standard error of the sample mean, $s_{\overline{x}}$. $s_{\overline{x}}$ may be calculated from the sample standard deviation, sd, by using the following formula: $s_{\overline{x}} = \frac{sd}{n}$

where n is the size of the sample.

Although $s_{\overline{x}}$ gives some idea of the variability of the sample mean, it is difficult to interpret without further assumptions or information. A more informative approach is to construct a confidence interval on the sample mean. A confidence interval is an interval about which we can make specific statements when appropriate assumptions are met. For example, suppose a random sample of cases reviewed resulted in a sample mean of 100 days in training and a standard error of 3 days. A 95% confidence interval would be (94.1, 105.9). This interval may be interpreted to mean that the probability is .95 that the interval (94.1, 105.9) contains the true population mean. Alternatively, we might say that if many samples of size 25 were drawn from the population and such intervals calculated, on the average, 95% of such intervals would include the true population mean. Confidence intervals for sample proportions may be interpreted in a similar manner.

More details regarding the construction of confidence intervals on means and proportions may be found elsewhere in this manual (see Unit Four, Module Five). Further information may be found in McCall's **Fundamental Statistics for Psychology** (1975) or Conover's **Practical Nonparametric Statistics** (1971).

Problem — assessing relationships between variables. It is sometimes the case that a program evaluator wishes to investigate the relationship between two variables. That is, two different variables are measured on each client in a sample resulting in a group of **matched pairs**. For example, this data might consist of pre-training and post-training measures of vocational functioning for each client. It might also be hypothesized that time in training and earnings at closure may be related. For continuous interval level variables such as time in training and earnings at closure, an appropriate measure of the relationship between the two is the Pearson Product Moment Correlation. The computation of this statistic and details of its interpretations are discussed in Conover (1971) and McCall (1975).

Suppose vocational functioning is measured before and after training using a five point rating scale. Since this type of data is clearly not interval level, slightly different correlation measures may be appropriate. Two such measures are called **Kendall's Tau** and **Spearman's Rho.** Both of these indices of dependence are interpretable and widely accepted. Either may be used to determine whether earnings at closure and time in training are independent or if in fact there is a tendency for people who spend a longer time in training to have higher earnings at closure. Conover (1971) provides an adequate coverage of the computational and interpretation aspects of these two nonparametric techniques.

A third approach to problems of assessing relationships between variables is called **linear regression**. Linear regression is a sophisticated statistical technique which may **sometimes** be used to establish a method of predicting an outcome or **dependent** variable such as earnings at closure on the basis of some **independent** variable such as time in training or age at referral. When more than one independent variable is used to predict the dependent variable the technique is called **multiple linear regression**. It should be pointed out that rehabilitation researchers have been largely unsuccessful in utilizing this technique to predict the outcome resulting from rehabilitation services (Bolton, 1974).

Problem. A second type of problem may arise when the data is in the form of matched pairs. It is often desirable to investigate whether or not one variable tends to be smaller or larger than another variable measured on the same person. For example, are clients' earnings at referral generally smaller than clients' earnings at closure? Is the personal adjustment of rehabilitation clients better after training than before training?

There are many statistical techniques applicable to this type of question. Possibly the most widely used is the t-test for paired observations. This test is based on the assumption that both variables are normally distributed. This technique is discussed in more detail in many elementary statistics texts (e.g., McCall, 1975).

In many cases the assumption of normality is not warranted. When this occurs, it is usually more appropriate to select a nonparametric test. Such tests include the sign test, the Wilcoxon test and the Randomization test. One example of this, the sign test, is discussed in this unit. Computational aspects of the others are discussed in Conover, 1971.

The table which follows provides lists of several common statistical techniques by type of hypothesis tested or question answered. The techniques listed are divided into parametric and nonparametric methods. The parametric methods require the assumption that the data come from normally distributed populations while the nonparametric methods require no assumption of normality. It is important to realize that there are some assumptions which must be met in order to test hypotheses using nonparametric methods. These assumptions are technique specific and are adequately presented in Conover, 1971.

The techniques are also partitioned on the basis of the level of measurement for which they are appropriate — Interval, Ordinal, or Nominal for nonparametric methods and Interval/Ratio for the parametric methods. Four types of data are considered. The **single sample** consists of a group of observations from a single population. For example, observations may consist of the age for each rehabilitation client in a random sample from the specified population. **Paired observation** data consists of pairs of observations on individuals from a single population. Examples of this type of data are pretest — posttest data or time in service and earnings at closure for each client in a group of rehabilitation clients. Data for **two independent samples** consists of observations on the same variable for samples from two different populations. For example, it may be desirable to compare the time in referral status for clients served by two district offices. The data would consist of the time in referral status for a sample of clients for each office.

Example null hypotheses are given for most of the techniques listed. It should be emphasized that these are only examples and the list is not exhaustive. The precise hypothesis tested by a particular technique depends not only on the specific situation an evaluator has, but also on the assumptions he is willing to make about his data. Again the alternatives are discussed in Conover (1971).

120

Normality of Data	**Level of Data**	**Type of Data**	**Example Hypotheses**	**Common Techniques** [1]	
Non-normal Distributions	Interval	a. Single sample	H_0: The median is less than or equal to 0.	Wilcoxon test	NONPARAMETRIC METHODS
		b. Paired observations	H_0: The values of one variable tend to be smaller than the values of the other.	Wilcoxon test	
		c. Two indep. samples	H_0: The means of the two samples are equal	Randomization test (two independent samples)	
		d. Several samples	H_0: All populations have the same median.	Kruskal-Wallis test (only requires ordinal data)	
	Ordinal	a. Single sample	H_0: The sample comes from a given completely specified distribution.	Kolmogorov Smirnov (1 sample test)	
		b. Paired observations	H_0: The values of one variable tend to be larger than the values of the other.	Sign test	
		c. Two indep. samples	H_0: The two samples come from the same distribution.	Mann-Whitney test, Wald Wolfowitz test	
		d. Several samples	H_0: All populations have the same median.	Kruskal-Wallis test, Median test	
	Nominal	a. Single sample	H_0: The design comes from a given completely specified distribution. H_0: One characteristic of elements of the population is independent of a second characteristic.	χ^2 test for independence, goodness of fit	
		b. Paired observations	H_0: The probability of change from 'YES' to 'NO' is the same as the probability of change from 'NO' to 'YES.'	McNemar test	
		c. Two indep. samples	H_0: The probability of being in class 1 is the same for both populations.	χ^2 test of difference between groups	
		d. Several samples	H_0: The probabilities of being in class 1, class 2, ... class k respectively are the same for all populations.	χ^2 test for difference between groups	
Normal Distribution	Interval/ Ratio	a. Single sample	H_0: The mean of the population is zero.	Test for specific mean	PARAMETRIC METHODS
		b. Paired observations	H_0: The mean of one variable is greater than the mean of the other variable. H_0: The correlation between the two variables is 0.	Paired t-test, correlation, regression	
		c. Two indep. samples	H_0: The means of the two groups are equal.	t-test, analysis of variance	
		d. Several samples	H_0: The means of all groups are equal.	Analysis of variance	
		e. Single sample with multivariate data	H_0: EXTREMELY VARIED POSSIBILITIES	Factor analysis, cluster analysis, multiple regression, model fitting	
		f. Several samples with multivariate data	H_0: EXTREMELY VARIED POSSIBILITIES	Multivariate analysis of variance, discriminant analysis	

---

[1] Detailed descriptions of most of these techniques may be found in the following easy to read textbooks: Conover, W. J. **Practical Nonparametric Statistics**. John Wiley & Sons, New York, 1971. McCall, R.B. **Fundamental Statistics for Psychology**. Harcourt, Brace & World, Inc., New York, 1970.

References

Bolton, B. **Introduction to Rehabilitation Research**, Springfield, Ill.: Charles C. Thomas, 1974.

Conover, W. J. **Practical Nonparametric Statistics**, New York: John Wiley & Sons, Inc., 1971.

Ferguson, G. A. **Statistical Analysis in Psychology and Education** [**Third Edition**]. New York: McGraw-Hill Book Company, 1971.

McCall, R. B. **Fundamental Statistics for Psychology** [**Second Edition**]. New York: Harcourt, Brace & World, Inc., 1975.

Unit Four

Module Five

Confidence Interval on a Mean

Suppose an evaluation specialist, Charlie, wishes to estimate the mean number of miles from home to work for the successfully rehabilitated clients of a given supervisory area. Since he cannot interview every rehabilitated client in the district and is under strict financial restraints, he wishes to base his estimate on a sample of size n. Charlie knows that the estimate will not give him the exact population mean so he wants some idea of the precision of his estimate.

Charlie's associate, P.J., tells him that the **standard error** is the most commonly used measure of the variation of an estimate. "If the **sample mean**, $\bar{y}$, is used to estimate the population mean," he says, "the standard error of $\bar{y}$ is denoted by $s_{\bar{y}}$ and is computed by:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}(1-f)$$

where

$$sd = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

is the **sample standard deviation** of a single randomly selected observation and the quantity $1-f = (1 - \frac{n}{N})$ is the finite population correction. Consider the following sample of size 4, (3, 7, 4, 6), from a population of size 100. $s_{\bar{y}}$ may be computed in the following manner":

y_i	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
3	-2	4
7	2	4
4	-1	1
6	1	1
$y = 20$		10

$\bar{y} = 5$

$$sd = \sqrt{\frac{10}{4-1}} = 1.83$$

Thus $S_{\bar{y}} = \frac{183}{\sqrt{4}}(1-\frac{4}{100}) = (.92)(.96) = .88$

Charlie says, "So what!"

P.J. patiently explains that if Charlie were to take many samples of size four that he would observe a sample mean between $\bar{y}-2.35(s_{\bar{y}})$ and $y+2.35\ (s_{\bar{y}})$ or 5-2.35(.88) and 5+2.35(.88) 90% of the time. In the present example, this means that Charlie is 90% confident that the true mean number of miles between the home and work of the 100 clients is between 2.93 and 7.07. "That's just what I wanted to know," says Charlie, "Almost."

In the example above the interval (2.93,7.07) is called a 90% confidence interval on a mean. Further, details on the computation and usage of such confidence intervals are available in Walpole (1968), McCall (1975), and other elementary statistics texts.

Several general properties of estimators are worth mentioning at this time. A method of estimation is **consistent** if the estimate becomes exactly aqual to the population value when the entire population is sampled. A method of estimation is **unbiased** if the average value of the estimate, taken over all possible samples of the given size n, is exactly equal to the true population value. For simple random sampling, $\bar{y}$ is a consistent and unbiased estimate of the population mean $\bar{Y}$.

Sampling fraction and the f.p.c. When the **sampling fraction** f= n/N is **low** the finite population correction (f.p.c.) is closer to 1 and the size of the population has no appreciable effect on the standard error of the mean. That is, if the standard deviation is the same for two populations of size N = 100,000 and N = 10,000 respectively, a sample of 500 from the larger population gives almost as precise an estimate of the population mean as a sample of 500 from the smaller population even though the fraction of the population sampled is much smaller. Recall that the f.p.c. is 1 minus the sampling fraction. It should be noted that the effect of ignoring the f.p.c. is to **overestimate** the standard error of the estimate $\bar{y}$ and hence unnecessarily increase the size of the confidence interval.

Suppose we wish to estimate the ratio of two variables both of which vary from element to element. For example, in a caseload survey, the average number of interviews per client prior to status 10 may be estimated. Two things vary — caseload size and number of pre-10 interviews. The population parameter to be estimated is:

$$R = \frac{\text{Total \# of pre-10 interviews}}{\text{Total \# of active cases}} = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\bar{Y}}{\bar{X}}$$

Note that this type of estimation is common when the sampling unit is a group or cluster of elements (in this situation a caseload) and the interest is in the population mean per element (as opposed to per unit).

The sample estimate is $\hat{R} - \frac{\bar{y}}{\bar{x}}$

Note: $\hat{R}$ is an unbiased, consistent estimate of R.

Suppose a random sample of ten counselors were selected resulting in the following observations.

Counselor	Active Caseload Size (x_i)	Total pre-10 interviews (y_i)
1	110	300
2	200	460
3	60	300
4	100	290
5	150	150
6	220	265
7	125	300
8	140	290
9	190	500
10	185	420
	$\sum_{i=1}^{10} x_i = 1480$	$\sum_{i=1}^{10} y_i = 3275$

The estimate $\hat{R}$ is then given by

$$\hat{R} = \frac{\sum_{i=1}^{10} y_i}{\sum_{i=1}^{10} x_i} = \frac{3275}{1480} = 2.2$$

Thus, the average number of pre-10 interviews is around 2 per client.

Review

A. What happens to the finite population correction (f.p.c.) when the population becomes very large relative to the sample size?

B. Find the sample mean, sample standard deviation and standard error of the mean for the following sample from a population of size 400. How would these change if the population was infinitely large?

C. What is the effect of ignoring the f.p.c., i.e., assuming an infinite population, when in fact the population is finite?

Answers

a. The f.p.c. becomes very small.
b. The f.p.c. becomes 0 so the variance estimates become slightly larger.
c. Overestimation of the variances.

Unit Four

Module Six

Testing for Differences in Means

Evaluation researchers often find occasion to compare two populations with respect to their means on some variable. For example, Charlie was asked to compare the severely disabled clients to the nonseverely disabled clients with respect to their mean time in service statuses (i.e., statuses 14, 16 and 18). In this type of analysis, the time in service status is called the **dependent variable.** Other possible dependent variables on which the two populations might be compared include such things as earnings at closure, case service cost and time in referral status.

There are five common techniques which an evaluator might use to test the difference between means of populations. All five require that the sample(s) is (are) selected randomly. Further, all require the specification of a significance level α, the probability of rejecting the null hypotheses when in fact it is true. The assumptions which determine the situation in which each is an appropriate technique are given below.

t-test with equal variances

Assumptions:

1. Two independent random samples of size n_1 and n_2 respectively from two normally distributed populations.
2. The population variances σ_1^2 and σ_2^2 are unknown but equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma_2$.
3. The null hypothesis can be specified as H_0: The means of the two populations differ by a specified constant.

In the special case where this constant is 0, the null hypothesis is:

H_0: The means of the two populations are equal.

t-test with unequal variances

Assumptions:

1. Two independent random samples of size n_1 and n_2 respectively are selected from two normally distributed populations.
2. The population variances σ_1^2 and σ_2^2 are unknown but suspected to be unequal, i.e., $\sigma_1^2 \neq \sigma_2^2$.
3. The null hypothesis can be specified as H_0: The means of the two populations differ by a specified constant.
   When the specified constant is 0, the null hypothesis is:
   H_0: The means of the two populations are equal.

t-test for paired observations

Assumptions:

1. A random sample of size n is selected from a population and a **pair** of observations is made for each subject in the sample.
2. Both observations are normally distributed.
3. The null hypothesis can be specified as:
   H_0: The means of the two observations differ by a specified constant.
   When the specified constant is 0, the null hypothesis is:
   H_0: The means of the two observations are 0.

These three techniques for testing differences in means are all based on the **student t distribution.** The student distribution is really a family of distributions indexed by a parameter called the **degrees of freedom**. Hence, it is common usage to state that "the test statistic has a t distribution with k degrees of freedom." The degrees of freedom tell the researcher how to look up critical values for a hypothesis in a table of critical values for the t distribution.

Analysis of Variance

Assumptions:

1. Independent random samples of size $n_1, n_2, \dots n_k$ respectively are selected from each of k normally distributed populations.
2. Although the variances of the k populations are unknown they must be assumed equal. If the largest variance is more than 2 1/2 times as large as the smallest, this assumption is probably not justified and the technique is inappropriate.
3. The null hypothesis can be specified as H_0: The means of the k populations are all equal.

Analysis of variance is based on the normal distribution. Unlike the t distribution which is completely specified by a single parameter, the normal distribution requires two parameters. These two parameters are the familiar mean and standard deviation 1. Fortunately, in analysis of variance, this is the only distribution which must be considered.

The preceding four techniques are discussed in greater detail in many statistics texts. Walpole (1968) and McCall (1975) provide good coverage.

Kruskal-Wallis test

Assumptions:

1. Independent random samples of size $n_1 \dots n_k$ respectively are selected from each of k **non-normally** distributed populations.
2. The measurement scale is at least ordinal (see module on levels of measurement in Unit 3).
3. Either all k populations have identical distributions or the means of the population differ.
4. The null hypothesis may be stated H_0: The means of the k populations are identical.

The Kruskal-Wallis Test is called a non-parametric test since it does not require the assumption of data from the normal distribution (a parametric distribution). The critical values for the Kruskal-Wallis test must be found in a table especially for this test. Such tables may be found in any test on non-parametric statistics. Conover (1971) provides appropriate tables and an excellent discussion of the computation and use of the Kruskal-Wallis test.

There are many other possible tests for differences in population means. One is the **sign test** which is discussed in more detail in Module Seven. Others are given in the table of statistical techniques (Module 4). To use any of these techniques it is always advisable to consult a reputable text on statistical analysis. The texts appearing in the bibliography have been selected for the adequacy of their coverage and clarity of presentation.

The following pages give typical computer printouts of the statistical tests discussed above. If computer data processing is not available, Bitter and Goodyear (1974) include clear guidelines appropriate for desk calculator computation. Table 1 illustrates a computer printout of the Kruskal-Wallis test. The heading explains the printout. BMD refers to the computer program package, Biomedical Computer Programs, P35 refers to the specific computer program used. According to the heading there is one dependent or criterion variable, and four different groups. There are then two variables, the dependent variable (DEPVAR) and the independent variable (GRP). Reading across the DEPVAR line we see that all groups combined had a mean score of 2.60; standard deviation of 10.42 and a range of scores from 1 to 99. Under the heading "1 DEPVAR, Frequency" refers to the number (n) of persons in each group. The heading Rank Sum refers to the numbers that are tested according to the Kruskal-Wallis procedure. The actual test statistic (H) is 8.36, with three degrees of freedom (number of treatment groups (K) minus one, or **df**=K-1, **df**=4-1=3). This statistic is significant at .039. Thus we can reject the null hypothesis that the group means are identical or that the samples came from the same population and accept the alternative hypothesis. Thus we conclude that based on dependent variable scores, the groups are significantly different.

```
BMDP35 — NONPARAMETRIC STATISTICS
KRUSKAL-WALLIS H TEST

1 DEPENDENT VARIABLE    4 GROUPS

VARIABLE        MEAN          S.D.           MIN.          MAX.
No. Name
1 DEPVAR        2.6023        10.4237        1.000         99.000
2 GRP           1.8863        1.1979         1.000         4.000
```

overall mean for all groups

smallest observation over all groups

```
KRUSKAL-WALLIS ONE WAY ANALYSIS OF VARIANCE RESULTS

Variable                        1 DEPVAR              Rank
Group                           Frequency             Sum
No.      Name
1        STATE EVAL             51                    2076.0
2        RSA EVAL               13                     614.5
3        UNIV EVAL               7                     262.5
4        STATE ADMIN            17                     963.0
```

number of observations in Group 4

```
KRUSKAL-WALLIS TEST STATISTICS(H) 8.3624
LEVEL OF SIGNIFICANCE =  0.0391  ASSUMING CHI-SQUARE
DISTRIBUTION WITH 3 DEGREES OF FREEDOM
```

Since .0391 is less than .05, the four groups are significantly different.

Table 1

Illustration of Kruskal-Wallis Test

Table 2 is an example of a t-test between two independent random samples on a dependent variable. SPSS refers to the computer program "Statistical Package for the Social Sciences." The dependent variable is labeled POSTINFD,in this case a post treatment test completed by each group. There are 28 persons in each group. Group 2 has a mean posttest score of 26.02 and a standard deviation of 13.95. The standard error of the mean for Group 1 is 2.64. At an alpha level of .05, we can assume that the true population mean for the sample Group 1 will be between 23.4 and 28.6 with 95% confidence. That is, 95% confidence interval for the Group 1 mean is 26.01 ± 1.96 (2.638)(see Unit Four, Module Five). Under the heading "Pooled Variance Estimate" the t statistic equals 5.08 and with 54 degrees of freedom (df = N-2 or 56-2 = 54) is highly statistically significant. We can reject the null hypothesis of no difference between the means and accept the alternative hypothesis. Inspection of the group means reveals that Group 1's mean is significantly higher than Group 2's mean. If we suspect that the variances for the groups were unequal we would use the t statistic under the heading "Separate Variance Estimate." Note that in this analysis the degrees of freedom are reduced. Generally, the larger the degrees of freedom, the more likely we are to achieve statistical significance and reject the null hypothesis.

ONE DEPENDENT VARIABLE (POSTINFD), TWO GROUPS

```
---------------------------------T-TEST---------------------------------------
                           NUMBER                    STANDARD       STANDARD
VARIABLE                  OF CASES       MEAN        DEVIATION        ERROR
POSTINFD      GROUP 1        28         26.0179        13.959         2.638
              GROUP 2        28         11.2321         6.511         1.231
```

sample mean for Group 2

standard error of the mean for Group 1

```
                                    POOLED VARIANCE ESTIMATE
           F         2-TAIL      T         DEGREES OF       2-TAIL
         VALUE       PROB.     VALUE        FREEDOM         PROB.
Group 1
          4.60       0.00      5.08           54             0.00
Group 2
```

```
                SEPARATE VARIANCE ESTIMATE
                  T        DEGREES OF            2-TAIL
                VALUE       FREEDOM              PROB.
Group 1
Group 2          5.08        38.22                0.00
```

Test statistic

Since 0.00 is less than .05, the difference between the two group means is significant.

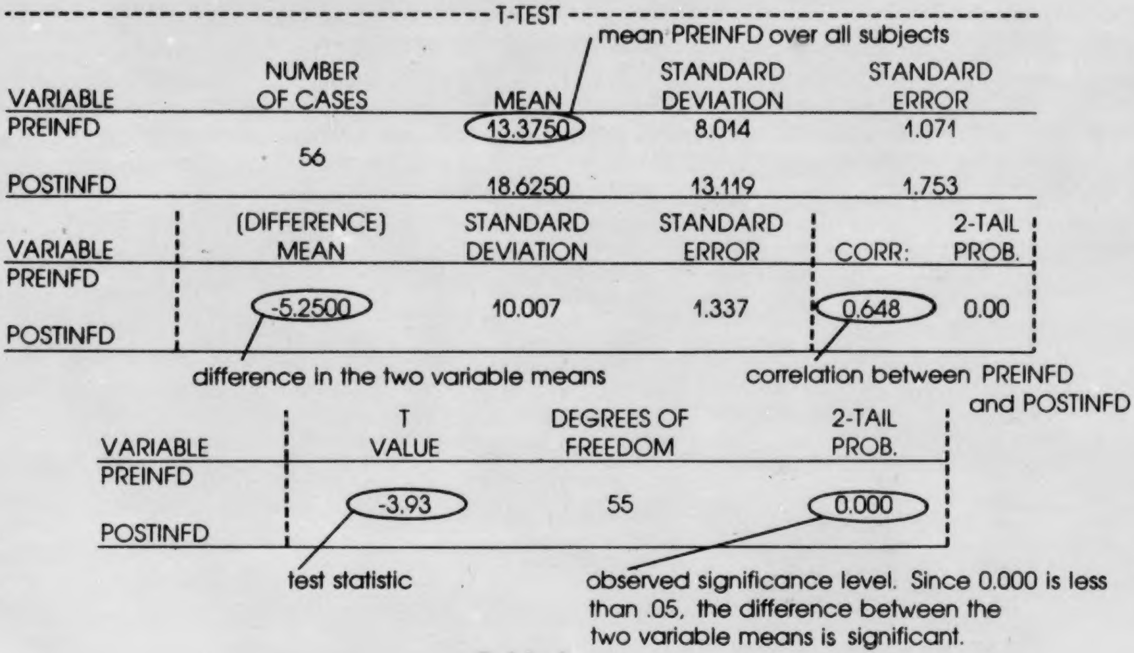
Table 2

Illustration of t-test comparisons between group means when dependent variable scores are independent

Sometimes it is necessary to test mean differences between paired or correlated scores. This is the case in testing for "gain" and "post" measures; it is likely that individual characteristics will systematically affect both pre and post scores. Table 3 presents an example of a paired t-test analysis. In Table 3 note that there are two mean scores, a "pre" and a "post" on one dependent variable INFD for one group of 56 persons. The difference between the two mean scores is -5.25. The correlation between individual pre and post scores is .64 which is statistically significant (.0001), indicating a nonchance relationship. The t-test statistic (-3.93) with 55 degrees of freedom (N-1 or 56-1 = 55) is highly significant. We can reject the null hypothesis that the means are equal and conclude that the post mean is significantly higher than the pre mean.

SSPS PAIRED T-TEST

TWO VARIABLES (PREINFD, POSTINFD), ONE GROUP



T-TEST

mean PREINFD over all subjects

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	STANDARD ERROR
PREINFD	56	13.3750	8.014	1.071
POSTINFD		18.6250	13.119	1.753

VARIABLE	(DIFFERENCE) MEAN	STANDARD DEVIATION	STANDARD ERROR	CORR:	2-TAIL PROB.
PREINFD POSTINFD	-5.2500	10.007	1.337	0.648	0.00

difference in the two variable means

correlation between PREINFD and POSTINFD

VARIABLE	T VALUE	DEGREES OF FREEDOM	2-TAIL PROB.
PREINFD POSTINFD	-3.93	55	0.000

test statistic

observed significance level. Since 0.000 is less than .05, the difference between the two variable means is significant.

Table 3

Illustration of t-test comparison between group means when dependent variable scores are paired

When one wishes to test three or more means or treatments the t-test analysis is generally an inappropriate procedure. Analysis of variance (ANOVA) is a technique by which means of general populations can simultaneously be compared. Table 4 presents the simplest case of analysis of variance, the one way ANOVA. The basic principle underlying one way ANOVA is to partition the total variation in the observed dependent variable into two identifiable sources — within groups variation and between groups variation. For example, under the heading "Source" in Table 4, Between groups and Within groups refer to the two mutually exclusive sources explaining the variation in the dependent variable. Between group variance may be thought of as the variation between the different group **means**, or the differences due to treatment. Within group variation refers to individual variability within the respective groups. The quantities under the heading "Mean Squares" represent estimates of the Between and Within groups variation respectively. If a large part of the variability in the dependent variable is due to the treatment, then it is to be expected that the Between group's mean square is large relative to the Within group's mean square. When the ratio of Between group's mean square to Within group's mean square (the F ratio) is large enough we say that the ratio is significant and that the group means are more different than could be expected by chance. In this case we reject the null hypothesis of equal means and conclude that the means are in fact different.

SPSS-ONE WAY ANOVA

```
PA/VR Card 16PF                                   10/10/75           Page 53

---------------------------------- One Way ----------------------------------

Variable 007

                              Analysis of Variance

Source            D.F.    Sum of Squares     Mean Squares       F        F Prob
Between Groups      2        16.2312            8.1156        1.716      0.190
Within Groups      45       212.7688            4.7282
Total              47       229.0000

    Group          Count        Mean          SD          Min.          Max.
     01              27        6.6296       0.3964        2.000         9.000
     02               9        6.4444       0.6894        4.000        10.000
     03              12        5.25         0.7191        3.000        10.000

Total                48        2.2073       0.3186        2.000        10.000

95 Pct.       Conf For Mean
Group
  01         5.8149 to 7.4444

  02         4.8546 to 8.0343
  03         3.6674 to 6.8326

Test for Homogeneity of Variance
  Bartlett - Box F =                                 0.304, P =  0.686
  Cochrans C = Max. Variance/ Sum C variances = 0.421, P =  0.370
```

degrees of freedom

F ratio test statistic

observed significance level. Since .19 is larger than .05, F is not significant.

standard deviations for Group 1

sample mean for Group 2

smallest value observed

mean of Group 1 will fall between 5.8 and 7.4, 95% of the time

P larger than .05 indicates assumption of homogeneity met

Table 4

Illustration of a One-Way Analysis of Variance

In Table 4, it is easy to see that the means for the three groups are 6.63, 6.44 and 5.25 respectively. The F-ratio of Between to Within groups mean squares is 1.72 which can happen 19 times out of 100 by chance (F prob = 0.190). This probability is too large for us to ignore, hence the null hypothesis of equal means must be **accepted** for the example in Table 4.

Review

1. Describe the difference between the assumptions of the three types of t-test.

2. Describe the difference between the assumptions of the Kruskal-Wallis test and Analysis of Variance.

3. Find the following quantities from Table 3.
   a. The smallest observation from group 3. ________
   b. The mean for group 2. ________
   c. The between group mean square. ________
   d. A 95% confidence for the mean of group 3. ________ to ________

References

Bitter, J.A., & Goodyear, D.L. **Rehabilitation evaluation: some application guidelines**. Greeley, Colorado: Regional Rehabilitation Research Institute, University of Northern Colorado, July, 1974.

Conover, W.J. **Practical Nonparametric Statistics.** New York: John Wiley & Sons, Inc., 1971.

McCall, R.B. **Fundamental Statistics for Psychology (Second Edition).** New York: Harcourt, Brace & World, Inc., 1975.

Walpole, R.E. **Introduction to Statistics.** New York: The MacMillan Company, 1968.

Unit Four

Module Seven

The Sign Test

Charlie was asked to assess the effectiveness of an inservice training program on interpersonal skills for counselors. As part of the training program, the level of each counselor's skill was rated from 1 to 5 before and after the training. Charlie was elated that finally he could use the paired t-test he learned back in Statistics 101. With this test, he reasoned, he could test whether or not the mean rating after training was in fact higher than before training. While compiling the data, however, P.J. noticed two things: First, neither the pretest nor the posttest appeared to be normally distributed and the sample size of 15 was not large enough to warrant any normal approximation; second, the data was clearly ordinal rather than interval level. P.J., somewhat enjoying the role of wet blanket, proceeded to rain on Charlie's parade. "You can't use the t-test, Charlie." After patiently reiterating the latest discoveries, P.J. smugly waited for Charlie's predictable response. "Ok, what procedure **should** we use?"

By now it should be obvious that P.J. is smart enough not to get carried away by small triumphs. Realizing that he may have to do the calculations for this evaluation, he carefully chooses the simplest of the possible techniques — the sign test.

"The sign test is simple computationally, requires only ordinal level data and may be used to test whether or not the post-test ratings tend to be larger than the pretest ratings."

"Sounds great to me," Charlie replies. "Let's get on it."

The data were compiled on the 15 counselors and for each pair of scores a plus (+)was assigned if the post rating was greater than the pretest rating and a minus (−)if the post rating was less than the pre rating. If the two ratings were identical then the observation was eliminated from further consideration and the sample size reduced accordingly. The results of the data collection follow:

Counselor	Pretest rating	Posttest rating	
1	2	3	+
2	3	5	+
3	1	4	+
4	4	5	+
5	2	4	+
6	2	4	+
7	2	4	+
8	3	5	+
9	1	1	0
10	4	3	−
11	5	4	−
12	4	3	−
13	2	3	+
14	2	3	+
15	3	2	−

To test the null hypothesis that the training had no effect; i.e., the counselors were just as likely to regress as they were to improve, against the alternative that the counselors tended to improve, P.J. found that the following steps must be performed.

1. Select the level of significance, α, at which the hypothesis is to be tested. Charlie and P.J. set α at .05.
2. Tabulate the data assigning a + or − to each pair depending on whether the post rating was larger or smaller than the pre rating.

3. Count the number of +'s and use this value as the test statistic.
4. Consult a table of the Binomial distribution to find the **critical value** for the test statistic. (Recall that the null hypothesis is rejected in favor of the alternative when the value of the test statistic equals or exceeds the critical value. The critical value depends on the significance level α.)
5. Reject or accept the null hypothesis according to whether the test statistic exceeds or is less than the critical value.

P.J. has selected a significance level of $\alpha=.05$ and has counted 10 + signs and one tie so the test statistic is 10 and sample size is 14. The table below containing critical values for sign test indicates that the critical value for $\alpha=.05$ and $n=14$ is 11. Since the test statistic is 10, which is less than the critical value, the null hypothesis is accepted and they conclude that the training had no significant effect.

Critical Values* for the One-sided Sign Test

	$\alpha = .05$	$\alpha = .01$
n= 5	5	5
6	6	6
7	6	7
8	7	8
9	8	9
10	9	10
11	9	10
12	10	11
13	10	12
14	11	12
15	12	13
16	12	14
17	13	14
18	13	15
19	14	15
20	15	16

*Test Statistic must equal or exceed critical value to reject H_0.

Review

1. A group of 35 rehabilitation clients were asked to indicate whether they were satisfied or not satisfied with their rehabilitation sevices. Following this assessment, they were shown a film giving information about the funding, policies, and rationale of the rehabilitation agency. After the film, the clients were again asked to indicate whether they were satisfied or dissatisfied. The data is summarized in the following table.

		After	
		Satisfied	Dissatisfied
Before	Satisfied	10	3
	Dissatisfied	17	5

This may be translated into 17 +'s, 3 −'s, and 15 ties. For $\alpha = .01$, $n = 20$, what is the appropriate critical value?________. Is the null hypothesis of no film effect accepted or rejected?______

2. The sign test is used with data in the form of __________.
   a. a random sample
   b. matched pairs
   c. a long list
3. The following page is a worksheet for computing a two tailed sign test. Complete this worksheet using data from question 1.

136

Sign Test Worksheet
(two tailed test)

1. Enter the number of untied pairs, n= ____.
2. Enter the total number of +'s, T= ____.
3. Enter the level of significance desired, α= ____.
4. Enter the table of critical values below and find the critical region corresponding to your choice of n and α:
   $T \leqq$ __________ $T \geqq$ __________

Selected Critial Values for Sign Test
(two tailed test)

	$\alpha = .05$	$\alpha = .01$
n	Reject H_0 if	Reject H_0 if
10	$T \leqq 1$ or $T \geqq 8$	$T = 0$ or $T \geqq 9$
12	$T \leqq 2$ or $T \geqq 9$	$T \leqq 1$ or $T \geqq 10$
14	$T \leqq 2$ or $T \geqq 10$	$T \leqq 1$ or $T \geqq 12$
16	$T \leqq 3$ or $T \geqq 12$	$T \leqq 2$ or $T \geqq 13$
18	$T \leqq 4$ or $T \geqq 13$	$T \leqq 3$ or $T \geqq 14$
20	$T \leqq 5$ or $T \geqq 14$	$T \leqq 3$ or $T \geqq 16$

5. Decision:

(a) If the observed value of T (from step 2) satisfies either of the inequalities in step 4, then **reject H_0** and conclude that +'s and -'s have different probabilities of occurrence. That is, either the +'s are more probable or the -'s are more probable.

(b) If the observed value of T (from step 2) **does not** satisfy either of the inequalities in step 4, then **accept H_0** and conclude that +'s and -'s are equally likely to occur.

Unit Four

Module Eight

Chi-Square Tests

Chi-square tests are one of the most popular ways to analyze classification or categorical data. Nearly all evaluation researchers will, at some time or another, run across the Chi-squared test for independence. This technique is used to test the null hypothesis that the rows and columns of the table represent independent classification schemes. For example, if a sample of 110 rehabilitation clients was classified according to disability type and according to religious affiliation, a researcher could test the hypothesis that the two classification schemes are independent through consideration of a table similar to the following.

Table 1
Contingency Table for Disability and Religious Affiliation

	Visual	Hearing	Amputation	Orthopedic	Psychological	Other	Row Margin Totals
Protestant	2	6	3	5	10	9	35
Catholic	1	2	1	6	12	10	32
Jewish	3	4	1	3	8	10	29
Other	2	1	1	1	1	7	14
Column Margin Totals	8	13	6	15	32	36	110

Table 1 illustrates the observed frequencies of disability by religious affiliation in the sample of 110 rehabilitation clients. For example, in the first cell (upper left hand corner of Table 1) two visually impaired clients are Protestant. Continuing down the column, one visually impaired is Catholic, three are Jewish, etc. The Chi-square statistic is computed by the following formula:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

where: χ^2 = Chi Square

Σ = means to sum

O = observed frequencies

E = expected frequencies

$(\quad)^2$ = square this quantity

Table 1 provides the observed frequencies. However, the expected frequencies must be computed. First add the observed across rows and down columns. These sums are the **marginal totals.** For the 'Protestant' **row**, the marginal total is 35; for the 'visual' **column** the marginal total is eight. Expected frequencies for each cell are calculated by multiplying the marginal row and column totals associated with each cell and dividing by the total sample size (N = 110). Thus for the visually impaired Protestants the expected frequency is $\frac{35 \times 8}{110} = 2.54$, for the visually impaired

Protestants the expected frequency is $\frac{35 \times 8}{110}$=2.54, for the Jewish visually impaired the expected frequency is $\frac{8 \times 32}{110}$=2.11 and so on (See Table 2). Placing the observed and expected values in tabular form facilitates computation of Chi-square (illustrated by Table 3).

Table 2
Expected and Observed Frequencies for Disability by Religious Affiliation Contingency Table, Expected Frequencies in Parentheses

	Visual	Hearing	Amputation	Orthopedic	Psychological	Other	
Protestant	2 (2.54)	6 (4.14)	3 (1.91)	5 (4.77)	10 (10.18)	9 (11.45)	35
Catholic	1 (2.33)	2 (3.78)	1 (1.75)	6 (4.36)	12 (9.31)	10 (10.47)	32
Jewish	3 (2.11)	4 (3.43)	1 (1.58)	3 (3.95)	8 (8.44)	10 (9.49)	29
Other	2 (1.02)	1 (1.65)	1 (.76)	1 (1.91)	2 (4.07)	7 (4.58)	14
	8	13	6	15	32	36	110

Table 3
Chi-Square Calculated from Data in Table 2

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
2	2.54	-.54	.29	.11
6	4.14	2.14	4.27	1.03
3	1.91	1.09	1.18	1.08
5	4.77	.23	.05	.01
10	10.18	-.18	.03	.002
9	11.45	2.45	6.00	.52
1	2.33	-1.33	1.76	.75
2	3.78	-1.78	3.16	.84
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
7	4.58	2.42	5.85	1.27

$\Sigma = 110$

$\Sigma = 10.44$

$X^2 = 10.44$

Following computation of Chi square it is necessary to compare the test statistic to a critical value found in the table of critical values for Chi squared distribution. To make this comparison the evaluator must do two things. First, the **level of significance,** α (the probability of rejecting the null hypothesis of independence when in fact it is true),must be specified. Second, the critical value for a Chi squared random variable with (r-1) (c-1) **degrees of freedom** must be found. Since "**chi-square**" refers to a whole family of distributions the appropriate one must be identified by a parameter called **degrees of freedom** (df). It so happens that the correct Chi squared distribution is determined by the number of rows (r) and columns (c) of the contingency table. That is, **df** for Table 1 equals four rows minus one (3) times six columns minus one (5) or 15. Tables for the family of Chi squared distributions can be found in most elementary statistics textbooks. For a significance level of $\alpha = .05$, the critical value for a Chi squared random variable with (3)(5) = 15 degrees of freedom is 25. The researcher must reject the null hypothesis if the test statistic **exceeds** the critical value. In the current example, 10.44 is less than 25, hence the null hypothesis is accepted and it is concluded that Disability and Religious Affiliation are independent classifications. They are apparently not related.

This Chi squared test for independence is based on a single sample of N observations. Each observation is classified into one of r row categories and c column categories resulting in an r x c table of frequencies. This same table and test statistic may be used to test an entirely different hypothesis when the sample is selected in a slightly different manner. Suppose that four random samples of size 35, 32, 29 and 14 were drawn from active rehabilitation clients with religious affiliations of Protestant, Catholic, Jewish, and Other respectively. If each observation was classified as to disability, the resulting table of frequencies would appear precisely as in Table 1. It would then be reasonable to test the null hypothesis that each of the four populations had the same proportions of the various disabilities. That is, it is possible to test that the distribution of disabilities is the same for all four populations (religious affiliations.) This is called a test of independence and compares the result to the same critical value of a chi-squared distribution. Hence, if Tables 1 and 2 were the result of this type of sampling the researcher would conclude that all four populations were identical relative to their disability composition.

Finally, a simple and frequently occurring type of contingency analysis is the fourfold or 2 x 2 classification table. Suppose an evaluator wants to determine if graduation from a rehabilitation center is associated with subsequent employment. For this analysis the x^2 test for independence can be computed without calculating expected values.

The analysis makes use of the table below, where marginal frequencies are represented by the letter outside the block, and the observed frequencies are represented by the letters within the block. Figure 1 represents 100 rehabilitation clients classified as to graduation and employment status.

A	B	A + B
C	D	C+ D
A+C	B+D	N

		Graduated		
		yes	no	
employed	yes	20	40	60
	no	25	15	40
		45	55	

Figure 1

Graduation by Employment Status Contingency Table

The Chi square test statistic is computed using the formula

$$\chi^{2'} = \frac{N(AD-BC)^2}{(A+B)\ (C+B)\ (A+C)\ (B+D)}$$

where the letters correspond to those in Figure 1. The test statistic for the current example is given by:

$$\chi^2 = \frac{100\ (20 \times 15 - 40 \times 25)^2}{60 \times 40 \times 45 \times 55} = 8.25$$

(Example from Ferguson, 1971, p. 186)

With degrees of freedom of (r-1) (c-1) or df = (2-1) (2-1) = 1, the χ^2 of'8.25 exceeds the critical value of 6.64 (from table of critical values) at .01 level of significance. Graduation from the rehabilitation center is significantly related to employment. Note that the above formula can also be applied to the data in the preceeding discussion on using χ^2 as a test of differences in proportions.

A Short Cut to Computing χ^2 in Contingency Tables

In large r x c contingency tables the following prodecure which deletes the necessity of computing O-E may be used.

1. Classify each observation by the row and column variable and construct the contingency table. Let O be the **observed** frequency in the i_{th} row and j_{th} column.
2. Compute the row total for each row of the table.
3. Compute the column total for each column total of the table.
4. Compute the **expected** cell frequency for each cell of the table. E_{ij} denotes the expected cell frequency for the cell in row i and column j and is computed using the following formula:

$$E_{ij} = \frac{R_i\ C_j}{N}$$

where
R_i is the row total for row i;

C_j is the column total for column j; and,

N is the total sample size.

5. Compute the test statistic χ^2 using the following formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N,$$

where r is the number of rows and c is the number of columns. This formula may be computed by following the steps listed below,

a. Beginning with the upper left hand cell (called cell (1, 1) since it is in the row 1 and column 1,) square the observed frequency, O_{11}, in that cell and divide the resultant number by the expected frequency E_{11} to obtain the quantity $\frac{O_{11}^2}{E_{11}}$.

b. Move to the next cell on the right, cell (1, 2), and compute a similar quantity $\frac{O_{12}^2}{E_{12}}$.

c. Continue computation of the quantities $\frac{O_{ij}^2}{E_{ij}}$ for all cells in the table.

d. Sum all of the quantities just computed to obtain $\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}}$.

e. Subtract N, the total sample size, from the sum obtained in step d to produce the desired chi-squared test statistic.

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N$$

For the frequencies in Table 1, expected frequencies are given (in parentheses) in Table 2. The ratios derived in steps 5a, 5b, and 5c $\left(\frac{O_{ij}^2}{E_{ij}}\right)$ are given in Table 3 below.

Table 3

Intermediate Computations for Disability By Religious Affiliation Contingency Table

	Visual	Hearing	Amputation	Orthopedic	Psychological	Other	
Protestant	1.57	8.69	4.71	5.24	9.82	7.07	
Catholic	0.43	1.05	0.57	8.26	15.47	9.55	
Jewish	4.27	4.66	0.63	2.28	7.58	10.54	
Other	3.92	0.61	1.32	0.52	0.98	10.70	
							120.44

From Table 3, $\sum_{i=1}^{4} \sum_{j=1}^{6} \frac{O_{ij}^2}{E_{ij}} = 120.44$,

hence the test statistic x^2 is given by $x^2 = 120.44 - 110 = 10.44$.

Review

1. Verify several entries in Tables 2 and 3.

2. For the following table, test the hypothesis that the row and column variables are independent.
   Note: The $\alpha = .05$ critical value of a chi-squared random variable with 2 degrees of freedom is 5.991.

	Black	White	Other
Male	14	27	32
Female	5	19	3

3. For the following table, test the hypothesis that the proportions are different for the two populations. Note: The $\alpha = .05$ critical value for a chi-squared random variable with 1 degree of freedom is 3.841.

	Severely Disabled	Non-Severely Disabled
District Office 1	63	12
District Office 2	31	84

Chi-Square Test for Differences in Probabilities
Worksheet for 2 x 2

1. For each of the two samples, enter the number of observations Class 1 and Class 2 in the "observed data" table below.

Observed Data

	Class 1	Class 2	Total
Sample 1	O_{11} = ____________	O_{12} = ____________	R_1 = ______
Sample 2	O_{21} = ____________	O_{22} = ____________	R_2 = ______
Total	C_1 = ____________	C_2 = ____________	N= ______ (grand total)

2. Compute the total number of observations in class 1 (C_1) and Class 2 (C_2) and enter into the appropriate blanks.
3. Compute the total number of observations in sample 1 (R_1) and sample 2 (R_2) and enter into the appropriate blanks.
4. Compute the grand total (N) and enter into the appropriate blank.
5. Compute the expected cell frequencies and enter into the table below. Use the following four steps for your computations.

   a. Compute $E_{11} = \frac{R_1 C_1}{N}$ = ______________.

   b. Compute $E_{12} = \frac{R_1 C_2}{N}$ = ______________.

   c. Compute $E_{21} = \frac{R_2 C_1}{N}$ = ______________.

   d. Compute $E_{22} = \frac{R_2 C_2}{N}$ = ______________.

   e. Enter E_{11}, E_{12}, E_{21}, and E_{22} into the expected frequencies table.

Expected Frequencies

E_{11} = ______________	E_{12} = ______________
E_{21} = ______________	E_{22} = ______________

6. Compute the test statistic χ^2 using the following formula and information from steps 1 and 5.

$$\chi^2 = \frac{O_{11}^2}{E_{11}} + \frac{O_{12}^2}{E_{12}} + \frac{O_{21}^2}{E_{21}} + \frac{O_{22}^2}{E_{22}} - N$$

χ^2 = ______ + ______ + ______ + ______ - ______

χ^2 = ________ .

7. Select the desired level of significance, α = ________ .

8. Enter the table of critical values for a 2 x 2 chi square test below and find the critical value corresponding to your choice of α and enter it here: CV = ________ .

Table of Critical Values for 2 x 2 Chi Square Test

	α = .05	.01	.001
Critical Value (CV)	3.841	6.635	10.83

9. Decision:

(a) If the observed value of χ^2 from step 6 exceeds the critical value (CV) from step 8 reject H_0 and conclude that population 1 and population 2 differ with respect to the relative proportions in Class 1 and 2.

(b) If the observed value of χ^2 does not exceed the critical value (CV) accept H_0 and conclude that the two populations are identical with respect to the relative proportions in Class 1 and Class 2.

References

Ferguson, G.A. **Statistical Analysis in Psychology and Education** [Third Edition]. New York:McGraw Hill Book Company, 1971.

Unit Four

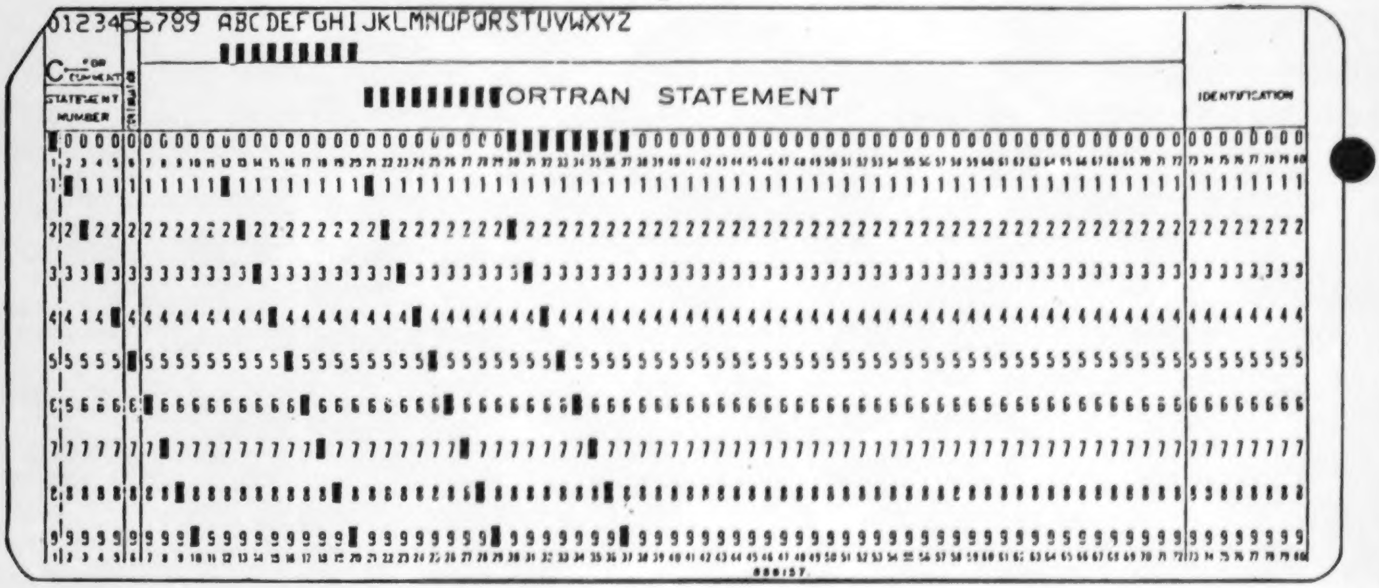
Module Nine

The Language of Data Processers

Program Evaluators often have the occasion to communicate with the data processing staff of their agency. Unfortunately, program evaluators and data processing personnel don't always speak the same language. In this module, five useful data processing concepts are discussed in language a program evaluator can understand.

The computer card. The standard computer card is made of cardboard and is approximately 3½″ by 7½″. It is marked off into 80 columns and 12 rows. By punching holes in one or more of the rows in a given column (using a keypunch machine), the column may represent a number, a letter or one of several symbols such as the ampersand (&) or comma (,). Obviously, the number 1 is represented by a punch in the row labeled 1 on the card. Likewise, the number 2 is represented by a punch in the row labeled 2. Letters and symbols are represented by multiple punches, that is, punches in several rows of a given column. The figure below shows the appropriate punches for digits 0 through 9 and the alphabet a through z. Note how each letter of the alphabet is represented by a multiple punch.

Figure 1



The format. When data such as the R-300 information is keypunched onto computer cards, each piece of information appears in a given set of columns on the card. For example, a client's age at referral is entered in columns 28 and 29 while the code for his disability at referral is punched in columns 31, 32, and 33. The term **format** is used to mean a statement indicating the location on the computer card where each piece of information is keypunched. In more technical use, the term format refers to a coded statement which tells the computer how many variables are to be read from the card, where on the card each variable appears, whether the information consists of letters and symbols or numbers, and where decimal points should be placed. If a given "record" for an individual contains more information than can be punched on a single card, several cards may be used and the format tells the computers how many cards comprise a record. The importance of maintaining accurate up-to-date files of formats for corresponding card decks cannot be overemphasized. If formats are lost, entire data sets may be rendered useless or require re-punching.

Tapes. All modern computing systems have facilities for using magnetic tape for data storage and manipulation. The use of tape storage is most efficient when data sets are large. The importance of accurate formats for tapes is even more important than for cards since the content of a tape cannot been "seen." It should be noted that while a card is limited to 80 characters, a tape may be considered as a series of cards each containing many more characters (The maximum number varies across different computers). Further documentation is also required for tapes. For example, the **tape density** must be known. That is, how many bits of information per inch (BPI) does the tape have. Most tapes are now 1600 BPI. It is also necessary to know whether the tape is **labeled or unlabeled** and whether it has **even or odd parity**. A labeled tape has information written on it regarding how much data is on the tape and how it is stored (not just on the tape reel but on the **tape**). **Parity** refers to a system through which the computer checks the tape for errors due to flaws in the physical tape itself, e.g., dirt or stretching. The designation of even or odd parity determines how this check is to be made.

Canned Programs. Some of the most useful data processing tools are the "canned programs" which are currently available. These are usually sophisticated computer programs designed to analyze, describe, or manipulate your data. Familiarity with the salient aspects and capabilities of one or two of these program packages can greatly facilitate the data analysis work of the program evaluator. Further, such familiarity can be a great advantage in communicating with data processing personnel. Some of the more useful program packages are:

SAS – Statistical Analysis System
SAS Institute Inc.
P.O. Box 10066
Raleigh, North Carolina 27605
SPSS – Statistical Package for the Social Sciences
Stanford University, Stanford, California
BMD – Biomedical Computer Programs
University of California
Berkeley, California

Programming Languages. It sometimes happens that a commercial program package is not available to perform the precise analysis or data manipulations required for a given evaluation study. In this case it may be necessary for the data processing staff to write a program using one of the many programming languages. Some of these are FORTRAN, COBOL, ALGOL, PL1, BASIC and RPG. Using these languages, a competent programmer can write routines to do almost any data analysis or manipulation imaginable. Most modern computing facilities support one or more of these languages, and it can be a great advantage for the program evaluator to understand the increased capabilities these languages give the data processing staff.

Review

1. How many columns are on the standard computer card?__________

2. The letter A is represented by how many punches?______________

3. The __________tells the computer which columns on the card are to be read.

4. The magnetic tape is most efficient when used for data sets which are______.

   a. small
   b. large
   c. unintelligible
   d. lost

5. BPI refers to__________.

   a. Being Persistently Irate
   b. Big Project Inducements
   c. tape density
   d. bits per inch

6. Two computing tools the program evaluator may use are ______.

   a. canned programs and programming languages
   b. odd and even parity
   c. Visa and MasterCharge
   d. a pencil and an eraser

Unit Four

Module Ten

Considerations in Presentation

At the conclusion of any program evaluation or evaluative research effort, the evaluator eventually finds it necessary to present the results of his study to administrators, management officials and other evaluators. The techniques used in the presentation of evaluative research findings are important since they affect the extent to which results are understood and hence the extent to which they are utilized. Following are several suggestions which can improve the clarity and utility of evaluative research.

Graphs. A picture can often clarify a lot of garbled verbiage, and graphs can often summarize a large volume of exposition. Bar graphs, pie graphs and frequency polygons can be effectively used to present results in a clear and concise manner. The graphs below are self-explanatory.

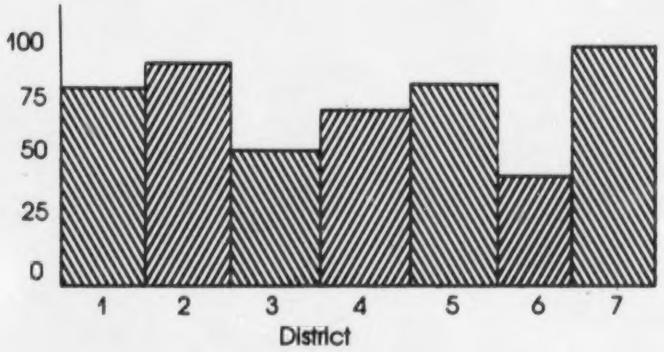


Figure 1
Percentage of Referrals Closed 08 by District

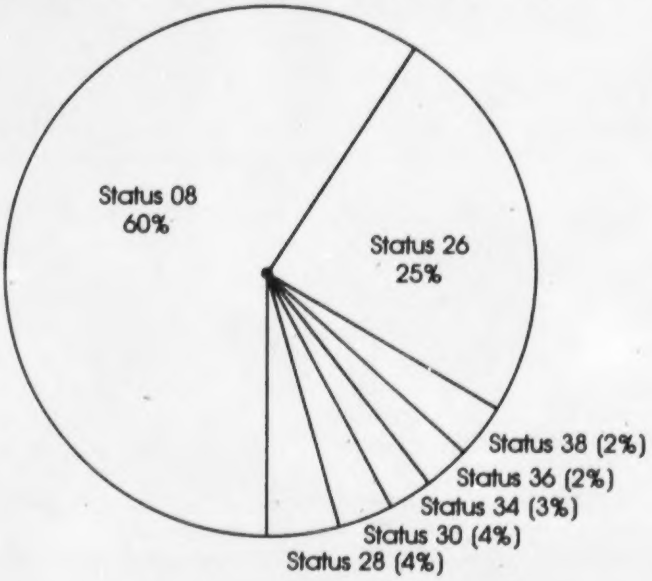


Figure 2
Client Closure Types for FY 1977

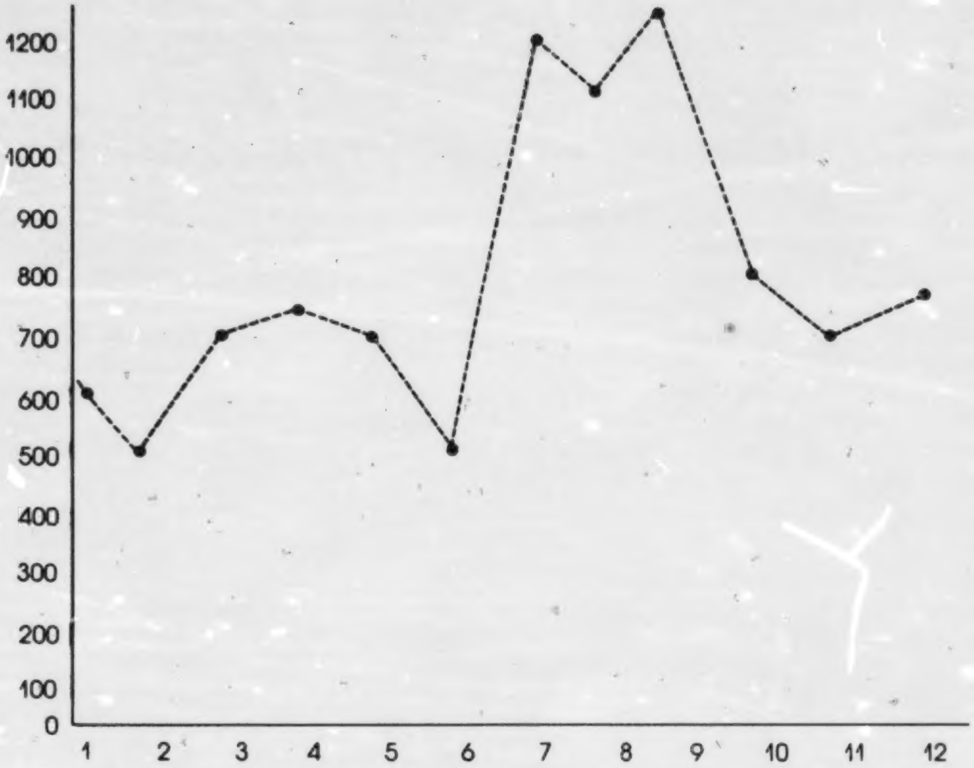


Figure 3
Client Referrals by Month for FY 1977
in District Office A

Means. Program evaluators often find it convenient to summarize large quantities of data by reporting means. There are two types of means which are commonly reported — sample means and population means. No matter which type of mean is being reported it is always important to include some measure of the variability of the population. When reporting a sample mean, the sample standard deviation and sample size are necessary to assess the precision of the reported mean by allowing computation of the **standard error of the mean**. Since we know a sample mean will not have the same value as the true population mean, some assessment of how close it is likely to be is necessary.

For example, suppose Charlie, estimating the average time in status 02, selects a sample of size 4 and computes a sample mean of 15 days with a standard deviation of 5. Suppose P.J. selects a sample of size 100 and reports a mean of 12 days with a standard deviation of 10. The standard error of the mean reported by Charlie is:

$$se = \frac{5}{\sqrt{4}} = \frac{5}{2} = 2.5$$

The standard error of the mean reported by P.J. is: $se = \frac{10}{\sqrt{100}} = \frac{10}{10} = 1.$

Even though the standard deviation is larger for the second sample the standard error of the mean is smaller, hence, the second mean is more precise.

If a population mean is reported, that is, information has been collected on every element of the population, then the **exact value** of the population mean is known; there is no reason for assessing the precision of an estimate. However, reporting the standard deviation of the population gives readers available information regarding the likelihood that a given individual will be near the mean. Clearly, in populations with larger standard deviations there is less chance of an individual being within say two points of the mean than in populations with small standard deviations.

Significant Findings and Implications for Practice. Administrators, management officials and evaluators all find it very helpful to have the significant findings and implications of a study specifically pointed out. This not only makes it easier to determine the applicability of a study to a particular problem but also increases the probability that the research will be utilized.

Documentation. The final suggestion for the presentation of evaluative research results concerns replicability. If adequate records and notations are kept by the researcher and presented in his final report, then subsequent studies may validate the original findings. Further, idiosyncrasies of the population studied, missing or edited data, sampling bias, and data collection problems **may** limit the generalizability of the results and therefore should be carefully documented. Additionally, it is often the case that precise methodology cannot be recalled from memory after the passage of a few years, and written documentation often becomes the saving grace of the evaluator who must for some reason recall what he did three years ago.

Review

1. Illustrate the use of the bar graph, the pie graph, and a frequency polygon.

2. Why should standard deviations and sample sizes be reported whenever a mean is given.

3. List 5 reasons for adequate documentation.

Epilogue

Our purpose in writing this text was to introduce the reader to issues and concepts of evaluation research in vocational rehabilitation. We have specifically avoided detailed reference to the current Federal evaluation "standards," concentrating instead on a skill building approach. We believe that the understanding and application of any regulatory evaluation standards require a working knowledge of basic program evaluation principles. We have also assumed that most persons responsible for evaluation in rehabilitation agencies and facilities are most likely to have had little formal training in evaluation methodology. Although there are of course, exemplary individuals and evaluation units within the vocational rehabilitation system, this training package was targeted toward those program evaluation specialists in state vocational rehabilitation agencies who have had little research training.

Experience as a practitioner, a service provider in the system being evaluated, is an advantage for any program evaluator. Yet the demands of program evaluation require that the evaluator possess special knowledge and objectivity. We can define the ideal evaluator as an individual who has mastered statistics, system engineering, research methodology, economics, and public relations, has the skills of a computer expert, business-manager, clinical psychologist, and group facilitator as well as experience as a successful author, a practitioner, and is independently wealthy! Obviously very few persons engaged in program evaluation would meet all of the above criteria.

While few of us can expect to master all of the skills and knowledge of the ideal program evaluator, we can develop basic skills and the knowledge of how to tap available resources to supplement those skills. Within this text we sought to introduce the reader to program evaluation in general and specifically to the methodology underlying evaluation research. Unit I included a general introduction to program evaluation, how program goals and objectives are related to an evaluation effort, issues inherent in conducting program evaluation, and a rationale for using an evaluation research approach to program evaluation. Unit II introduced sampling concepts. Sampling methodology is very important when resources are small and the populations to be investigated are large. When the evaluator is concerned with estimating manageable case load size, estimating client satisfaction with agency services and in drawing generalizations from experimental programs, it is important that correct sampling procedures are followed. Unit III reviewed measurement concepts. Virtually every type of objective program evaluation requires that entities be quantified so that they may be categorized, analyzed, and interpreted. The principles of measurement underlie what is often euphemistically referred to as "data elements" in current program evaluation literature. Unit IV, "Data Analysis and Research Design," described specific techniques for drawing conclusions from evaluation data and some of the issues regarding interpretation of evaluation data.

By now it should be apparent that program evaluation is a complex undertaking which consists of interlinked activities each of which place limitations on the other. For example, consider Standard Number One of the original Federal program evaluation standards for vocational rehabilitation. This standard stated in effect that one goal of evaluation was to insure that the rehabilitation program served its target population and to insure that services were provided in an equitable manner. This evaluation goal relates to the larger program goal, the effectiveness of rehabilitation services. Note however, that while the evaluation goal, or standard, is clear in its intent, it does not clearly specify the operations the evaluator should use in assessing goal achievement. For example, what do "services" and "equitable" mean? How will we define "target population"? In order to determine if the rehabilitation program is equitably serving the "target population," we need to define that population by considering if there is a "priority" target population, e.g., the severely disabled, define what we mean by disabled, severely disabled, (occupationally handicapped and/or physically impaired per se, functionally disabled, etc.[1]) and devise a method to measure severity of disability. If we can agree that the target population of vocational rehabilitation is "disabled persons," it follows that the population

---

[1] Physical impairment means the medical condition, disability refers to functional limitations resulting from the impairment, and handicapped suggests environmental barriers to functioning.

154

will differ in **degree** of disability, hence, we must consider such measurement problems as scaling, reliability and validity. Assuming we have an adequate measure of disability, we now try to estimate, using sampling procedures, the population of disabled persons and compare, using data analysis techniques, the total population of persons who might be served by the rehabilitation program to those actually served. The way in which we choose our definitions or criteria places limits on how we will measure those criteria, which in turn limits our sampling and data analysis procedures.

Evaluation Guidelines

The above example illustrates three points: the interrelationship of program evaluation data elements, the complexity of program evaluation and the increasing involvement of the Federal Government in setting national human service evaluation standards. We agree with Franklin and Thrasher(1976) who have suggested that the Federal government will evolve minimum evaluation standards, an activity already well under way in vocational rehabilitation, a process for monitoring programs and eventual specification of what constitutes effective program participant outcome. We consider these regulations as minimum monitoring procedures basic to any large organizational enterprise. And we realize that vocational rehabilitation evaluation specialists are quite concerned, and understandably so, with development of program wide evaluation standards.We do, however, see the establishment of standards as a means to an end, not an end in themselves. Any program specific evaluation standards or regulations must be flexible and constantly changing because of shifts in legislative mandates and consequent changes in program priorities. Even when standards are clearly specified, the evaluation operations needed to fullfill evaluation requirements are often unclear. Hence, on-line agency and facility evaluators need to develop a basic understanding of program evaluation techniques and methodology. Finally, we would be remiss not to point out that accountability is but one function of program evaluation. In our view the real power of program evaluation is as a function of program improvement. Whereas national evaluation can standardize evaluation data over the entire state-Federal rehabilitation program those data are generally not discreet enough to address specific state, district, agency or facility concerns.

Resources for Evaluators

By now the reader is probably asking, "What do I do now?" "What's next?" One approach is to size up your immediate working situation. It is important to determine how much administrative support, both moral and financial, is available for program evaluation activities. Does, for example, the evaluation unit have line or staff responsibilities? Is the evaluation unit a separate entity on the organizational chart, and is it responsible to an administrator high in the organization's hierarchy? An experimental scale, the "Organization Readiness to Accept Program Evaluation Questionnaire," available from the Program Evaluation Resource Center, 501 Park Ave., Minneapolis, Minnesota, 55415, may help you judge your organization's responsiveness to program evaluation.

Assuming that there is organizational support for program evaluation, evaluators may wish to upgrade their skills. Training resources include local universities and colleges, most of which offer courses in statistics and experimental design. With the continuing importance of evaluation, the Federal government spent over $130 million on evaluation activities in 1975; various agencies such as the National Institute of Mental Health and the National Institute of Education are sponsoring formal university based evaluation training programs. One such training program has been developed at Northwestern University, Evanston, Illinois. That program trains evaluation reseachers on a full time basis. Several rehabilitation counseling programs have incorporated a course or courses in program evaluation and the Rehabilitation Services Administration has been active in sponsoring national and regional conferences regarding program evaluation.

Other resources for the evaluator include the rehabilitation research and training centers and rehabilitation research institutes. These facilities are dispersed regionally across the nation and have faculty skilled in research and program evaluation methodology. Generally, faculty are available for consultation and/or training at no expense to the rehabilitation agencies within their region. These facilities also periodically publish evaluation related materials.

Besides relevant resources texts which may be found in the bibliography, there are several periodicals which deal with program evaluation. The ICD Rehabilitation and Research Center, 340 East 24th Street, New York, New York, 10010, publishes an information exchange newsletter, "Program Evaluation in Vocational Rehabilitation." Items listed in the newsletter may be borrowed from ICD. Similarly, the National Clearing House of Rehabilitation Materials, Oklahoma State University, 115 USDA Building, Stillwater, Oklahoma 74074, publishes listings of program evaluation materials. The Center for the Study of Evaluation, 145 Moore Hall, UCLA, Los Angeles, California, 90024, publishes "Evaluation Comment" and the Evaluation Research Society, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts, 02115, publishes the "Evaluation Research Society Newsletter" and holds an annual convention. Membership is open to anyone engaged in evaluation research. Finally, at least two journals are devoted exclusively to program evaluation concerns. **Evaluation: A Forum for Human Service Decision Makers** is published by the aforementioned Program Evaluation Resource Center and is primarily targeted to evaluators working in mental health settings. **Evaluation Quarterly: A Journal of Applied Social Research** is published by Sage Publications, Inc., 275 South Beverly Drive, Beverly Hills, California, 90212.

Reference

Franklin, J.L., & Thrasher, J.H. **An introduction to program evaluation.** New York: John Wiley and Sons, 1976.

Evaluation Research: Annotated Bibliography

General Evaluative Research

Anderson, S.B., Ball, S., & Murphey, R.T. **Encyclopedia of Educational Evaluation.** San Francisco, California: Jossey-Bass, 1975. As suggested by the title this reference book truly is an encyclopedia covering evaluation from accountability to variance. Relevant terms are also listed by category.

Broskowski, A., Attkisson, C.C., Tuller, I.R., & Berk, J.H. **Human Services Program Evaluation: An Edited Bibliography of Key References.** Milwaukee, Wisconson: V-U Publishing Co., 1975. Contains references categorized by: Introductory-general, organizational theory, general research methodology, needs assessment, standards, etc.

Caro, F. (Ed.). **Readings in Evaluation Research.** New York: Russell Sage Foundation, 1971. This hardback contains thirty-one articles on evaluation of social programs. Well put together.

Chabotar, K.J., & Lad, L.J. **Evaluation Guidelines for Training Programs.** Lansing, Michigan: Michigan State University, Midwest Intergovernmental Training Committee, 1974. An easily read paperback covering the basics of evaluation research. Focused on evaluation inservice training. Sample evaluation measures are included in the appendix.

Franklin, J.L., & Thrasher, J.H. **An Introduction to Program Evaluation.** New York: John Wiley and Sons, 1976. This book is "intended to be a textbook for neophyte evaluators and practitioners as well as a reference book and source book for operating program evaluators." The text is easy reading and can serve as a first introduction in a program evaluation course.

Glass, G.V. (Ed.). **Evaluation Studies Review Annual.** Beverly Hills, California: Sage publications, 1976,**1**. The **Review Annual** yearly reprints timely evaluation articles from a variety of disciplines. The 1976 edition covers theory and method and studies in education, mental health, welfare, and crime.

Struening, E.L., & Guttentag, M. **Handbook of Evaluation Research.** Beverly Hills, California: Sage Publications, 1975, (Vol. I & II). A truly monumental work covering all phases of evaluation research. Each chapter is written by an evaluation expert.

Suchman, E.A. **Evaluation Research, Principles and Practices in Public Service and Social Action Programs.** New York: Russell Sage Foundation, 1967. This basic introductory text to program evaluation is fast becoming a classic. The book has three main sections covering conceptual, methodological and administrative approaches to evaluation.

Weiss, C.H. **Evaluation Research: Methods For Assessing Program Effectiveness.** Englewood Cliffs, New Jersey: Prentice-Hall, 1972. A well written and valuable overview of evaluation research. Chapters on "formulating the question and measuring the answer" and "utilization of evaluation results" are recommended. Contains an extensive bibliography.

General Measurement References

Anastasi, A. **Psychological Testing** (3rd ed.). New York: Macmillan, 1968. Contains information on methodological issues in testing, how to evaluate published tests and interpret tests. In addition, the text contains detailed information on various psychological tests and comprehensive bibliographies and where to obtain additional information.

Cronbach, L.J. **Essentials of Psychological Testing** (3rd ed.). New York: Harper and Row, 1970. This comprehensive textbook is often used in introductory courses on psychological testing. The text covers basic concepts as well as various types of tests such as tests of ability and typical performance.

Harris, C.W. (Ed.). **Problems in Measuring Change.** Madison, Wisconsin: University of Wisconsin Press, 1963. This book of readings contains several articles which have since become standards in the field on the problem of measuring participant change or "gain" following treatment.

Helmstadter, G.C. **Principles of Psychological Measurement.** New York: Appleton-Century-Crofts, 1964. A textbook designed for a basic course in psychological measurement. The book concentrates on measurement principles rather than instruments. Requires an understanding of basic statistics.

Lyman, H.B. **Test Scores and What They Mean.** Englewood Cliffs, New Jersey: Prentice-Hall, 1963. An extremely easy to read, non mathematical book on how to interpret test scores. Includes a glossary of terms, conversion table and summary of score characteristics.

Nunnally, J.C. **Psychometric Theory.** New York: McGraw-Hill, 1967. This text was written for graduate level psychology students planning on conducting psychological research. Considers the principles underlying psychological measurement.

Osgood, C.D., Suci, G.J., Tannenbaum, P.H. **The Measurement of Meaning.** Urbana, Illinois: University of Illinois Press, 1957. **The** book on the popular semantic differential.

Measurement Instruments

Bassett, P.T. (Chm.). **Measurement of Outcomes.** Denver, Colorado: First Institute on Rehabilitation Issues, West Virginia Rehabilitation Research and Training Center, 1974. A report of a conference considering issues in rehabilitation outcome measurement. Includes chapters on issues and legislation, review and analysis of current practice, and conclusions and recommendations. The appendix includes various types of measures used to assess client outcome.

Bonjean, C.M., Hill, R.J., & McLemore, S.D. **Sociological Measurement: An Inventory of Scales and Indices.** San Francisco, California: Chandler Publishing Co., 1967.

Botterbusch, K.F. **Tests and Measurements for Vocational Evaluators.** Menomonie, Wisconsin: Materials Development Center, University of Wisconsin-Stout, 1973. Contains descriptions (purpose, administration, content, scoring, materials required, appropriate groups, norms, reliability, validity, and sources of further information) on 33 psychometric tests. Covers some of the most popular achievement, personality, vocational, and aptitude tests for rehabilitation concern. Addresses of test publishers are listed.

Buros, O.K. **Mental Measurement Yearbooks.** Highland Park, New Jersey: The Gryphon Press. A series of seven mental measurement yearbooks, each with comprehensive reviews on published tests and extensive reference lists of how the tests have been used. The last yearbook (Seventh **Mental Measurement Yearbook**) consisted of two volumes with over 12,000 references. Also published by the same firm are: **Tests In Print,** a descriptive bibliography of 2467 commercially available tests; **Personality Tests and Reviews** and **Vocational Tests and Reviews.**

Comrey, A.L., Backer, T.E. & Glaser, E.M. **A Source Book For Mental Health Measures.** Los Angeles, California: Human Interaction Research Institute, 1973. This 462 page source book is truly "a taxonomy of data-collecting devices." Many little known, unknown and generally previously inaccessible instruments are referenced. Measuring devices from alcoholism questionnaires to vocational tests are cited.

Garwick, G., & Brintnall, J. Proceedings of the Second Goal Attainment Scaling Conference. Minneapolis, Minnesota: 1974. An overview of the basics on the popular goal attainment scaling method.

Hargreaves, W.A., Attkisson, C.C., Siegel, L.M., McIntyre, M.H. & Sorenson, J.E. (Ed.) **Resource Materials for Community Mental Health Program Evaluation: Part IV, Evaluating The Effectiveness of Services.**Rockville, Maryland: National Institute of Mental Health, 1975. Contains resource materials, some of which are adaptable to rehabilitation concerns. Examples of various mental health program effectiveness forms are given. Particularly recommended is the chapter on evaluating a training conference. Parts I, II and III deal with elements of program evaluation, needs and assessment planning and management planning and management information systems for Mental Health Centers.

Robinson, J.P., & Shaver, P.R. **Measures of Social Psychological Attitudes.** Institute for Social Research, University of Michigan, Ann Arbor, Michigan: 1970. Referenced are various "attitude" measures including life satisfaction scales, measures of self-esteem, values, alienation, general attitudes toward people, etc.

Zytowski, D.G., & Betz, E.L. Measurement in Counseling Research: a Review. **The Counseling Psychologist,**1972, **3**, 72-86. A compendium of published and unpublished measurement devices dealing with counseling concerns. Included are counselor selection instruments, counselor personality measures, counselor role measures, process measures and various measures completed by clients.

Vocational Rehabilitation Resources for Program Evaluation

Andrew, J.D., (Ed.). **Assessment of Rehabilitation Counselor Performance.** Stout, Wisconsin: First Institute on Rehabilitation Issues, University of Wisconsin, Research and Training Center, 1974. Methods of appraising rehabilitation counselor performance in relationship to agency requirements are covered. The emphasis is on the need and benefits of counselor appraisal systems. Examples of state agency appraisal systems are given.

Baptista, J., (Chm.). **Program Evaluation: A Beginning Statement.** Stillwater, Oklahoma: Oklahoma State University, Clearing House on Rehabilitation Materials, 1972. Designed as a "training manual" in program evaluation activities for rehabilitation agencies. This guideline covers such topics as "Why do program evaluation?" basic evaluation designs for agencies and the nature of program evaluation activities.

Bennett, E.C., & Weisinger, M. **Program Evaluation: A Resource Handbook for Vocational Rehabilitation.** New York: ICD Rehabilitation Center, 1974. A true resource handbook. Especially valuable in this 100 page paperback are the comprehensive bibliography, examples of relevant program evaluations and the compendium of index measures.

Berkowitz, M., Englander, V., Rubin, J., & Worrall, J.D. **An Evaluation of Policy Related Rehabilitation Research.** New York: Praeger, 1975. The results of a comprehensive study to "determine the methodological adequacy and policy utility" of rehabilitation research studies. See volume 20, number 1 of the **Rehabilitation Counseling Bulletin** for some experts' critical reactions to the study.

Bitter, J.A., & Goodyear, D.L. **Rehabilitation Evaluation: Some Application Guidelines.** Greely, Colorado: University of Northern Colorado, Regional Rehabilitation Research Institute, 1974. A workbook where ten examples of research design and statistical analysis are applied to rehabilitation evaluation problems. Each example contains a sample evaluation objective, criteria, methodology (measurement and statistical) and interpretation. The examples are clear and easily followed, no mathematical sophistication is required of the reader.

Bolton, B., (Ed.). **Handbook of Measurement and Evaluation in Rehabilitation.** Baltimore, Maryland: University Park Press, 1976. Consists of three sections, fundamentals of measurement, review of instruments and applications in rehabilitation. Chapters are written by well-known experts in the field. Of special interest to evaluators are the chapters on "measurement of client outcomes in rehabilitation" and "assessment of counselor performance."

Bolton, B. **Introduction to Rehabilitation Research.** Springfield, Illinois: Charles C. Thomas, 1974. An overview of statistical techniques and research design as applied to vocational rehabilitation concerns. Understandable by the novice, this book does not presume an extensive statistical or research background.

Bolton, B., (Ed.). **Readings in Rehabilitation Research: An Introduction to Methodology.** Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1972. Written to introduce rehabilitation research to practitioners, the author has included studies from the literature to illustrate "six basic approaches to research design" (description, measurement, prediction, nonexperimental, experimental, and factor analysis). Also included is a section on research utilization.

Cook, D.W. Guidelines for Conducting Client Satisfaction Studies. **Journal of Applied Rehabilitation Counseling,** 1977, **8**, 107-114.

Cooper, P., & Cook, D. **Program Evaluation in Vocational Rehabilitation: An Annotated Bibliography.** Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1974. This bibliography gives a representative sample of the program evaluation literature published by state rehabilitation agencies and other sources. The material is arranged to correspond to the RSA evaluation standards.

Greever, K.B., Leary, P.A., & Minton, E.B. **Program Evaluation: A Beginning Statement — Study Guide.** Institute, West Virginia: Rehabilitation Research and Training Center, 1973. An absolute rock bottom introduction to statistical manipulation of rehabilitation agency data (i.e., R-300 data). Includes using R-300 data and computing averages, using percentages, setting up a data deck for computer processing and reporting.

Greever, K.B., Minton, E.B., & Tseng, M.S. **Follow-up Studies of Rehabilitation Clients: A Step by Step Guide.** Institute, West Virginia: West Virginia Research and Training Center, 1974. A very basic guide to sampling and designing follow-up questionnaires.

Hamilton, L.W., & Muthard, J.E. **Research Utilization Specialists in Vocational Rehabilitation.** Gainsville, Florida: University of Florida, Rehabilitation Research Institute, 1975. A report on the ability of rehabilitation agencies to utilize research (and by implication evaluation) findings. Particularly valuable is the section on dissemination and utilization strategies.

Hills, W.G. Evaluating Vocational Rehabilitation Programs. **Monograph No. III.** Norman, Oklahoma: University of Oklahoma, Regional Rehabilitation Research Institute, 1973. Covers an overview of evaluation and rehabilitation and different measurement approaches to evaluation. Includes a "basic kit for measuring effectiveness."

McNickle, R.K., (Ed.). **Report of the Comprehensive Service Needs Study.** Washington, D.C.: The Urban Institute, 1975. This large volume contains an overview of the various concerns regarding the 1973 Rehabilitation Act.

Moriarty, J.B. **P.E. Puzzling You?** Institute, West Virginia: West Virginia Research and Training Center, 1976. A basic introduction, written in a light vein, to integrating and displaying agency evaluation data. The current RSA evaluation standards are covered in some detail.

Moriarty, J.B. **Profile Analysis Technique: Developing Standards for Self-Evaluation of VR Programs.** Institute, West Virginia: West Virginia Research and Training Center, no date. Consists of a rationale for using the profile analysis technique on data routinely collected by VR agencies, macro and micro aspects of program evaluation in rehabilitation and a chapter on the measurement of client outcomes in vocational rehabilitation.

Muthard, J.E., (Ed.). **Counselor and Program Evaluation Methods in Vocational Rehabilitation.** Gainesville, Florida: University of Florida, Rehabilitation Research Institute, 1977. The transcript of a training conference in vocational rehabilitation program evaluation. Included are short and readable descriptions of the profile analysis technique, an excellent section on program evaluation tips by Robert Walker and a section on methodological concerns in follow-up studies. The emphasis is on how-to-do-it.

Program Evaluation: A First Step. Chicago, Ill.: Commission on Accreditation of Rehabilitation Facilities, 1974. A short set of materials designed to encourage rehabilitation facilities to develop and install a program evaluation system. Examples of such a system and the CARF standards are included.

Rayman, I.C. Survey Techniques for Planning and Evaluation. In Robinault, I.P., & Weisinger, M., (Eds.)., **Program Planning and Evaluation: Selected Topics for Vocational Rehabilitation.** New York: ICD Rehabilitation and Research Center, Research Utilization Laboratory, 1975. An excellent review of concepts in sampling and survey design. Analysis of data and presentation of results are also discussed.

Reagles, K.W., Wright, G.N., & Butler, A.J. Correlates of Client Satisfaction in an Expanded Vocational Rehabilitation Program. **Wisconsin Studies in Vocational Rehabilitation: Monograph XII, Series 2.** Madison, Wisconsin: The University of Wisconsin, Regional Rehabilitation Research Institute, 1970. A study of client satisfaction with the Wood County project. The monograph illustrates various concepts in measurements and correlation analysis.

Robinault, I.P., & Weisinger, M., (Eds.). **Program Planning and Evaluation: Selected Topics for Vocational Rehabilitation.** New York: ICD Rehabilitation and Research Center, Research Utilization Laboratory, 1975.

Rubin, S.E., (Ed.). **Studies on the Evaluation of State Vocational Rehabilitation Agency Programs.** Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1975. This massive volume is the end product of RSA regional task force studies on program evaluation in vocational rehabilitation. Ten separate evaluation studies

addressing the then current evaluation "standards" are divided into three sections: procedures for estimating the disabled population, evaluating state rehabilitation agency process, and measuring client rehabilitation outcome.

Rubin, S.E., (Ed.). Issues in Program Evaluation Research in State Rehabilitation Agencies. **Conference Proceedings.** Fayetteville, Arkansas: University of Arkansas, Arkansas Rehabilitation Research and Training Center, 1974. A compendium of presentations including "Problems in measuring outcomes in rehabilitation research projects" and "Designing state vocational agency evaluation research."

Smits, J.J. A Model for Conducting Follow-up Studies in Medical Rehabilitation Settings. **Monograph No. 4.** Birmingham: University of Alabama, Medical Rehabilitation Research and Training Center, 1973. Points out the obstacles and pitfalls in conducting a follow-up study of medical inpatients. Includes the development of the rationale for the follow-up study, methodological procedures employed and analysis of the survey data.

Spaniol, L. A Model for Program Evaluation in Rehabilitation. **Monograph XIX, Series 3.** Madison, Wisconsin: Regional Rehabilitation Research Institute, 1975.

Weisinger, M., Robinault, I.P., & Bennett, E.C., (Eds). **Program Evaluation: Selected Readings.** New York: ICD Rehabilitation and Research Center, 1975. A paperback book of readings covering a wide variety of disciplines. It is divided into three sections: basic issues, models and approaches and techniques in evaluation. The third section contains techniques directly applicable to rehabilitation concerns.

Statistics and Design

Blank, S.S. **Descriptive Statistics.** New York: Appleton-Century-Crofts, 1968. A self-instructional programmed text covering the basics of descriptive statistics. Most readers can complete the 150 pages in less than five hours.

Budd, W.C., & Kelly, S.P. **Educational Research by Practitioners: An Elementary Casebook.** New York: Harper and Row, 1970. Easily understood examples of different types of research designs are given. The examples are all in an educational context.

Campbell, D.T., & Stanley, J.C. **Experimental and Quasi-Experimental Designs for Research.** Chicago, Ill.: Rand McNally, 1963. This thin book is **the** classic on research design. The evaluator will find the authors' treatment of quasi-expermental designs especially valuable.

Christensen, H.B. **Statistics Step by Step.** Boston, Massachusetts: Houghton Mifflin, 1977. A self-instructional textbook designed for independent study by students taking a first course in statistics. Each unit or module is divided into title, task, definitions, discussion, simplified discussion, examples and practice problems.

Cochran, W.G. **Sampling Techniques.** New York: John Wiley and Sons, Inc., 1963. A classic book in sampling theory. The book is designed for the more advanced researcher but provides an invaluable source book for the statistically oriented. Coverage of material is both concise and accurate.

Conover, W.J. **Practical Nonparametric Statistics.** New York: John Wiley & Sons, Inc., 1971. A book designed for applied researchers. It contains easily understood descriptions of most common non-parametric methods.

Dustin, D.S. **How Psychologists Do Research: The Example of Anxiety.** Englewood Cliffs, N.J.: Prentice-Hall, 1969. Descriptions of various types of research; exploratory studies, test development, sample survey and experiments are all illustrated by actual research studies. Easily understood by the novice.

Ferguson, G.A. **Statistical Analysis in Psychology and Education.** (4th ed.). New York: McGraw-Hill, 1976. An introductory text to applied statistics. The "emphasis is on the analysis and interpretation of data resulting from the conduct of experiments." Very useful as a cookbook type of reference.

Finney, D.J. **Experimental Design and Its Statistical Basis.** Chicago, Illinois: University of Chicago Press, 1955. Written for students in biology. This low cost paperback has a clear presentation of such relatively complex statistical designs as randomized blocks and Latin squares.

Glass, G.V., Willson, V.L., & Gottman, J. M. **Design and Analysis of Time-Series Experiments.** Boulder, Colorado: Colorado Associated University Press, 1975. An advanced reference text on interrupted time-series experiments.

Hartley, J.O. **The Design of Scientific Experiments**. Sigma Xi lecture, University of Arkansas, November, 1976.

Jones, R.H. (Ed.). **Methods and Techniques of Educational Research.** Danville, Illinois: The Interstate, 1973.

Kerlinger, F.N. **Foundations of Behavioral Research.** New York: Holt, Rinehart and Winston, 1964. This is probably one of the most comprehensive reference books on research for the program evaluator. The author stated that the book is aimed at graduate students "who have elementary backgrounds in psychology, statistics and measurement."

Kirk, R.E. **Experimental Design: Procedures for the Behavioral Sciences.** Belmont, California: Brooks/Cole, 1969. A text often used at the graduate level, contains examples of practically all the well known designs. The author uses a notation system which is explained in the text. Requires an understanding of college level algebra and statistical inference.

Kish, L. **Survey Sampling.** New York: John Wiley and Sons, 1965. An excellent reference book for sample survey methods. Although some of the book is definitely for more advanced readers, extensive sections are devoted to discussion of the practical aspects of survey sampling.

Langley, R. **Practical Statistics Simply Explained.** New York: Dover Publications, 1970. This low cost paperback purports to present in "plain English" basic concepts of inferential statistics. The emphasis is on significance tests for the non-mathematically minded.

McCall, R.B. **Fundamental Statistics for Psychology.** (2nd ed.). New York: Harcourt, Brace and Worlds, Inc., 1975. The book is designed for applied researchers in the behavioral sciences. It is written on an elementary level but includes adequate coverage of the basic data analysis techniques. Notation is simply explained and examples are plentiful.

McGuigan, F.J. **Experimental Psychology: A Methodological Approach.** (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall, 1968. This text is designed for use in a first level undergraduate course on experimental psychology. It is easily read and presents research design in a very understandable manner.

Moursund, J.P. **Evaluation: An Introduction to Research Design.** Monterey, California: Brooks, Cole, 1973. A very readable introduction to evaluation research. Besides the basics, includes information on directing a research project and the use of computers.

Neale,J.M., & Liebert,R.M.**Science and Behavior: An Introduction to Methods of Research.** Englewood Cliffs, New Jersey: Prentice-Hall, 1973. A basic text covering research concerns. The authors' approach is conceptual rather than statistical. The text appears to have been written for college sophomores. Recommended as a good and understandable overview of research methodology.

Robson, C. **Experimental Design and Statistics in Psychology.** Baltimore, Maryland: Penguin Books, 1973. A low cost paperback written at the introductory level. Contains worked examples for common statistical tests and a chart showing what test to use given the nature of measurement.

Siegel, S. **Nonparametric Statistics for the Behavioral Sciences.** New York: McGraw Hill, 1956. The intermediary text to nonparametric statistics. Especially valuable is the quick reference chart in the use of nonparametric tests given the level of measurement and sample size. Designed for applied researchers in the behavioral sciences, the book gives adequate coverage of most common nonparametric techniques. A good addition to the evaluator's library. "Cookbook" type instructions make the book useful to the occasional user.

Underwood, B.J. **Psychological Research.** New York: Appleton-Century-Crofts, 1957. A classic used in undergraduate behavioral science courses. The emphasis is on applying the methods of science to the study of behavior. The chapter on operational definitions is a classic in itself.

Willemsen, E.W. **Understanding Statistical Reasoning.** San Francisco, California: W.H. Freeman, 1974. The primary purpose of this paperback is to help persons evaluate research literature in the behavioral sciences. Assumes the reader has a knowledge of basic statistics.

Winer, B.J. **Statistical Principles in Experimental Design (2nd Edition).** New York: McGraw Hill, 1971. This text is considered one of the bibles on statistical principles underlying experimental design. It is a comprehensive reference and is generally used as text in advanced behavioral/biological graduate courses.